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:	*Publishing Paper*
Student ID number:	2311233
Word count:	4885
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.	

Assessment of fine-tuned open-source large language models on gold standard synthetic epilepsy clinical notes with focus on Seizure Frequency and Prescriptions

ABSTRACT

Background

Seizures impact people worldwide, leading to epilepsy. Extensive research has been conducted to understand seizure control, enhance clinician-patient interactions, and evaluate the effectiveness of specific medications. However, a lack of data has hindered these efforts. Clinical notes, maintained by clinicians contains seizure frequency and prescription details. Extracting this information using clinical Named Entity Recognition (NER) and Natural Language Processing (NLP) in healthcare could be a viable solution to address and meet the requirements for deriving epilepsy associated details. NLP has advanced significantly with the development of Large Language Models (LLMs). The focus of the study is to assess the performance of LLMs on gold standard synthetic epilepsy clinical letters.

Method

Exploratory data analysis (EDA) is performed to understand the data and structure the logic for data transformation in data pre-processing. The dataset is split into respective seizure frequency and prescription input-output pairs for model ingestion. Five open-source model candidates are selected for fine-tuning. Each model is trained for both seizure frequency and prescription separately. The model output is post-processed for valid JSON. Two separate validation methods are applied for assessing the quality of the generation.

Results

From the selected models, Llama 2 model generates output closer to the ground truth. The result has higher precision and lower recall. In prescription, the generation for keys of JSON files, Exact Match Ratio (1), Hamming Loss (0), Recall (1), Precision (1), F1 Measure (1); for values of JSON Exact Match Ratio (0), Hamming Loss (0.38), Recall (0.62), Precision (1), F1 Measure (0.76). In seizure frequency, the generation for keys of JSON files, Exact Match Ratio (1), Hamming Loss (0), Recall (1), Precision (1), F1 Measure (1); for values of JSON Exact Match Ratio (0), Hamming Loss (0.30), Recall (0.60), Precision (1), F1 Measure (0.74). The Sorensen dice coefficient for string in the text field for 50% of validation set < 0.5. In comparison to ExECT the performance is underwhelming.

Conclusion

Clinical NER in NLP is complex task. LLMs, based on transformer architecture, are used for entity extraction with focus on seizure frequency and prescription is a niche area. Data preprocessing and model fine-tuning are performed. The generated JSON files underwent post-processing and validation, but the model's performance suggests room for improvement. The results show underperformance compared to specialized models. Future improvements could make the model more scalable and accurate, with enhancements in datasets, modelling methodology, and validation techniques. Future advancements in NLP could further enhance the model's effectiveness.

INTRODUCTION

Seizures are known to affect global population which result in epilepsy (Falco-Walter, 2020). Epilepsy is a condition that affects general population without the consideration of age, race, geographical locations, and social class (Asadi-Pooya et al., 2023). There are studies that reported, the point prevalence of epilepsy as 9.37 per 1000 persons and the incidence rate as 42.68 per 100, 000 person years for United Kingdom (Wigglesworth et al., 2023).

The severity of epilepsy indicated and measured as seizure rate or seizure frequency is identified and reported by the clinician from the personal sessions with the patient. Seizure frequency is associated to the occurrence of seizure over a period by the patient (Baker et al., 1998; Choi et al., 2014; Viteva, 2014). The personal sessions are documented in the form of clinical notes. Outpatient clinical notes are documentation tools that records the patient condition and status for each session and interaction with clinician. The documentation helps in knowledge transfer from one clinician to another, online and offline tracking (Hanson et al., 2012; Soto et al., 2002; Wood, 2001). Seizure frequency is directly associated to outcomes like health life quality, depression, suicide, effectiveness of the drugs (Liu et al., 2020). Due to the importance of seizure frequency, researchers have focused on extracting the details associated that indicates seizure frequency. The extraction of seizure frequency is from clinical notes which are usually a form for unstructured data. The importance of seizure frequency is recently realized but however 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 (NLP) is a field in computer science that is formed by the intersection of artificial intelligence and linguistics domains that focus on trying to solve human language related problems like translation, context understanding, entity or label extraction that could be used to predict, forecast or derive insights from text-based documents (Jain et al., 2024; Katyan et al., 2024; Ramsay & Ramsay, 1987; Schank, 1986). Examples of application of NLP are chatbots, clinical text mining, clinical decision support (Jain et al., 2024; Katyan et al., 2024). The clinical documents mainly considered for NLP applications are radiology reports, biomedical reports, outpatient clinical notes (Demner-Fushman et al., 2009).

With the rise of transformers architecture and attention mechanism the field of NLP has found exponential progress and growth (Dean, 2022; Myers et al., 2024; Vaswani et al., 2017). The application of NLP to extract entities from clinical documents is termed as clinical Named Entity Recognition (NER). LLMs could be used as a tool for information extraction and NER. NER could be explained as extraction of semantic and syntactic meaning of the word with respect to the sentence and the context. This extraction happens on word level (Nasar et al., 2022; Reichenpfader et al., 2023). Due to privacy issues, the data is not publicly available for research purposes. The data is either siloed in protected vault and data centres with controlled access. However, for training LLMs large amount of data is required. In-order to circumvent this issue there are several techniques innovated. Some of the techniques are zero-shot learning, few-shot learning, prompt tuning, fine-tuning, retrieval augmented generation (RAG) (Brown et al., 2020; Fan et al., 2024; Liu et al., 2022; Naguib et al., 2024; Wei et al., 2022; Xiong et al., 2024).

BACKGROUND

Clinical NER in NLP was performed in pre-LLM era mainly using the techniques and tools from knowledge graph, machine learning, prompt tuning, deep learning, rule-based framework, statistics, and early implementations of transformers.

The techniques were used either individually or in combination to create customized framework. The combinations could be categorized as rule-based mix approach, deep learning mix approach, knowledge graph mix approach.

In the rule-based mix approach, the solution is designed by combining rule-based frameworks statistical techniques or machine learning. cTAKES was the first open source system for extracting information from unstructured data; specifically from electronic medical records that paved the way for clinical NER (Savova et al., 2010). Another example for an approach that was applied to extract specifically seizure frequency and prescriptions was based on the general architecture for text engineering (GATE) – ExECT using rule-based framework and statistical techniques to target nine categories of epilepsy clinical notes (Fonferko-Shadrach et al., 2019). General medical concept frameworks and implementation like UMLS, SNOMED, CT, and others were used as lookup databases for referencing standardized medical concepts, and retrieving concept mappings that were used as support elements for NER task.

In deep learning mix approach, the systems primarily used deep learning models that was combined with classical machine learning or statistical methods for clinical concept extraction (Chalapathy et al., 2016; Xie et al., 2022; Zhang et al., 2020). Word embeddings were popular applications to find and compare between two or more words or sentences and compute their cosine similarities in the vector space. Word embeddings could be weights of the customized shallow neural networks or from language models (Mikolov et al., 2013; Pennington et al., 2014; Zhu et al., 2018). Typically, these word embeddings are used for text conversion from string to numerical data type and then ingested to machine learning models or deep learning model to perform NER as a multi-class classification problem. The classes would represent the standardized codes from the medical concept frameworks or customized labels.

In knowledge graph mix approach, the systems were built with knowledge graph as an additional component to aid or support the model prediction thereby increasing the model performance. Knowledge graph was mixed with models like Bidirectional Encoder Representations from Transformers (BERT) and Conditional Random Field (CRF). The knowledge graph used in this setting is usually a biomedical knowledge graph (Harnoune et al., 2021).

In the LLMs era, the clinical NER was performed using LLMs enhanced with techniques like prompt tuning, prompt engineering, fine-tuning, RAG, knowledge graph and knowledge augmentation techniques (Y. Hu et al., 2024; Khan et al., 2024; Monajatipoor et al., 2024; Naguib et al., 2024; Xiong et al., 2024). The implementation consisted of large language models such Llama (Touvron et al., 2023), GPT (Xu, n.d.), Gemini (Gemini Team, Google, 2023) and others. The main approach found in the literature, to tackle clinical NER was to utilize technique likes prompt tuning, prompt engineering, fine-tuning, RAG, knowledge graph and knowledge augmentation techniques. The performance and results of the LLMs on the same dataset with fine-tuning and prompt related training were experimented and compared. It was observed that prompt tuning despite results in a non-persistent state of the model, it outperforms the persistent fine-tuned model. The probable explanation to this would be that the large language model typical consists of billions of parameters. Even after quantization and parameter efficient training (PEFT), there are a considerable number of parameters that needs to be learnt or adjusted. The size and quality of the dataset matters. Typically, healthcare dataset is smaller and imbalanced in nature. This might result into smaller weight updates. To fine tune the model, a considerable amount of dataset is needed. The corresponding cost is realized for training the model. However, in case of prompt tuning it only requires a smaller amount of upfront cost, no specific technical knowledge is required, cheaper and faster training time, and the ease to connect with external components like RAG gives an added advantage to this approach. Thus, prompt related training type models is more found in the literature than the actual model fine-tuning. However, in this study, realizing the drawbacks of fine tuning LLMs, it is considered as main method of solving the task of clinical NER with the focus to generate and extract seizure frequency and prescriptions from the clinical letters.

METHOD

The research development was performed as an iterative process (see Image 1.). There are 7 stages of development. The exploratory data analysis and data pre-processing logic were constructed gradually after observing and assessing the generated model output. The assessment was performed using samples from validation dataset by visual inspection.

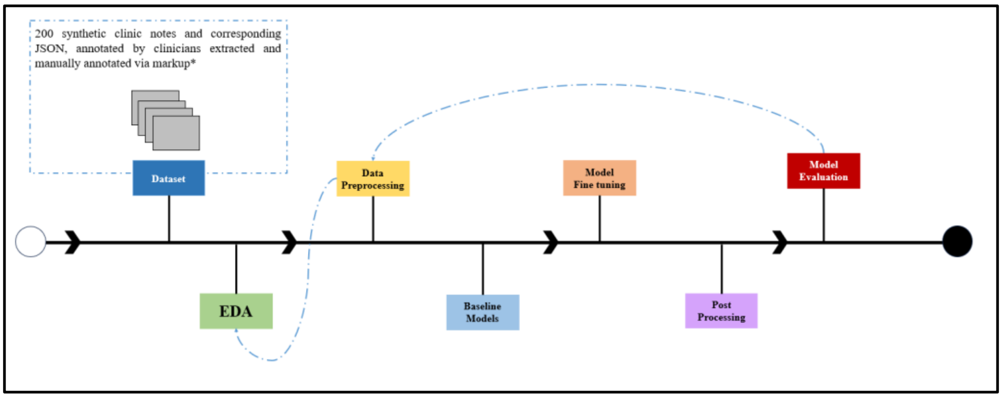

200 synthetic clinic notes and corresponding JSON, annotated by clinicians extracted and manually annotated via markup*
Dataset
EDA
Data Preprocessing
Baseline Models
Model Fine tuning
Post Processing
Model Evaluation

Image 1: Development Cycle

Dataset

An open source dataset was acquired from internet (Fonferko-Shadrach et al., 2023). The dataset consists of 200 synthetic clinical letters with corresponding annotations. The clinical files were stored as ".txt" files and the corresponding annotations as JSON format and ". json" files. Each clinical letter is reviewed and annotated by the expert clinician. Thus, the dataset could be considered as a gold standard. Despite the gold standard nature of the dataset, there were errors found in the dataset annotation. The errors are possible cases of mental fatigue encountered during the annotation process by the clinician. The JSON file consists of list of key-value structure. Each key-value structure represents the attributes and its corresponding values in reference to the clinical letters. For example, prescriptions would be tagged and represented in the key-value structure consisting of name of the prescription, dosage, start, and end index indicating the location of the text in the clinical letter.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed on the clinical letters that is the text files and JSON files separately. The purpose of the EDA was to understand and derive any insights that would aid in making decisions for data transformations in data pre-processing, model training and evaluation of the model results. There were three specific steps carried out in the EDA for clinical letters.

The first step in EDA for clinical letters, Image 2 represents the distribution of top 50 files and bottom 50 files, the count of all words and unique words. The word counts were calculated for the respective cases and sorted in descending order. This step helped in identifying the outliers at the file level. There were files with extremely large number of words in comparison to the general scenario. Average size of the documents with unique set of words and all words could be determined. This information helps tokenization of the input to the model.

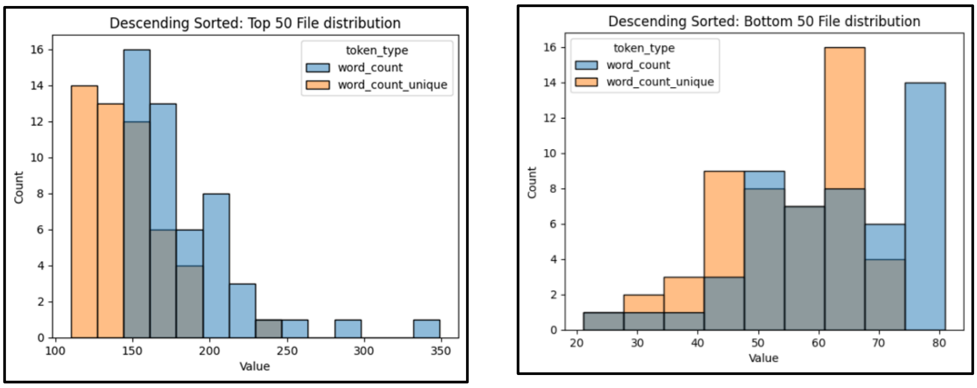


Image 2: Distribution of word count in each clinical letter

The second step in EDA for clinical letters, Image 3 represents the distribution of all the words and unique words count in the entire clinical letters collection. This helped to identify the outlier at the dataset level. Also, helped to identify the average length of the string that would be used as configuration setting for model input tokenizer.

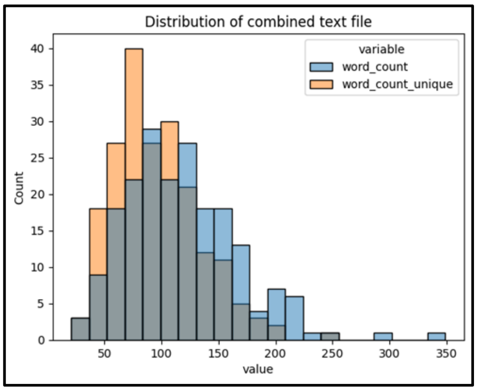


Image 3: Distribution of word count in entire clinical letter dataset

The third step in EDA for clinical letters, Image 4 represents the importance of the words. The importance of the word is understood by two approaches, "Word Cloud" and "TF-IDF". While the word cloud represents the word with the highest frequency with larger font size (Jin, 2017), the term frequency – inverse document frequency (TF- IDF) a term weighted scheme in information retrieval that weights the words based on the occurrence of it across the document collection (Aizawa, 2003). The TF-IDF chart from Image 4 represents the word importance in the clinical text in the descending order. It is observed that "seizure" word is most frequent and most important.

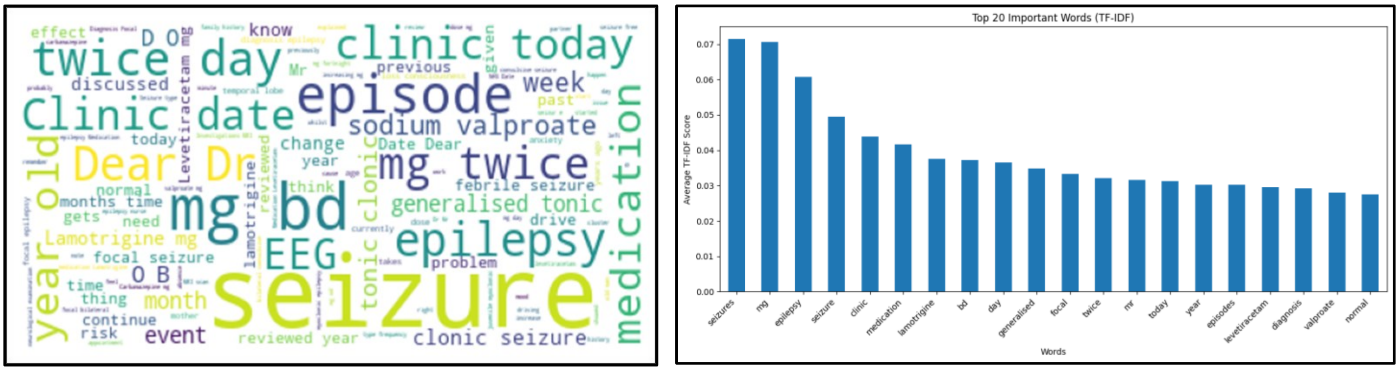


Image 4: Word Importance in the clinical text

Data Preprocessing

Data preprocessing was performed on the clinical letters and JSON files to prepare the data for model training. The first step is to understand the data in-order to design the solution and the logic for data transformation. During the data checks for JSON file, it was found that there were files where seizure frequency and prescription were both absent, in some files only prescriptions were present and in others only seizure frequency were present. There were more than one prescription and seizure frequency entries in the JSON (see Image 5). For prescriptions there were 3 and seizure frequency 25 unique data dictionary structure. This indicates that the target files have different JSON structures. These insights helped in decision making for data transformations. It was finalized to design separate input-output pair for seizure frequency and prescription. For each clinical letter there would be a target file that will only contain either the seizure frequency or the prescription JSON. It became essential that the input prompt which consists of the expected output structure for each of the sample needs to be customized corresponding to the JSON file. The JSON files needed to be cleaned and simplified in-order to make the JSON generation task easy for the model.

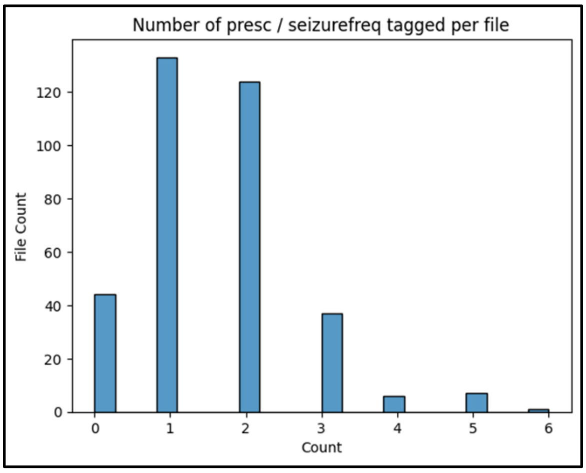


Image 5: Number of entities per file

The second step in data preprocessing was to clean and simplify the dataset. For the clinical letters, additional new line spaces, tab spaces were removed. Characters existing due to the UTF-8 encoding errors were resolved. For the JSON files, anything except the seizure frequency and prescription related data were filtered out. Certain key-values like the CUI-ID, CUI phrase, hyphens from the text field in the JSON file and entity key were removed.

The third step consisted of designing the prompt template input for the model. Since it was observed that there were multiple patterns of the JSON structure to ensure the correct schema is capture a dynamic schema extraction logic was implemented. The prompt contained sample output schema for model guidance of JSON generation.

In the fourth step, each input prompt designed in the previous step was mapped to a JSON file containing exactly one JSON dictionary. The mapping was then stored in ".jsonl" file, which is basically a JSON format but implemented using newline characters to separate JSON values.

Baseline Models

There were 5 models considered for fine-tuning. The model candidates were considered based on the criteria that, model should be open source, it contains small or medium size parameter count to be able to fit in GPU memory, and the optional condition that the model could be pretrained with medical data. The selected model candidate names from hugging face are, Pennlaine/Mistral-7B-v02-Entity-Extraction, NousResearch/Llama-2-7b-chat-hf, NousResearch/Hermes-3-Llama-3.1-8B, mistralai/Mistral-7B-Instruct-v0.2 and google/gemma-2-2b. The models are gated models; thus, they require special access and authorization to download and use the model for individual purpose under an open-source license agreement. The dataset was split to form two separate subsets of data that contained specifically input-output pairing of seizure frequency and prescription. Thus, the model was customized for each dataset, that is, a model for seizure frequency and a separate model for prescription were fine-tuned. So, in total, there were 10 models that were trained parallelly on the GPU.

Model Training Process

Model training process consists of 4 phases (see Image 6).

In the phase 1, model setup, the model of choice is downloaded along with its tokenizer from hugging face. Each model has its own unique tokenizer. The model is then quantized using post-training quantization technique (Kim et al., n.d.) and then the quantized model is combined with parameter efficient fine tuning (PEFT) using low rank adaptation technique (LoRA) (E. J. Hu et al., 2021) for memory efficient and faster fine tuning during the training phase. Finally, a supervised fine tuning trainer (*Supervised Fine-Tuning Trainer*, n.d.) is configured for dynamic input prompt formatting.

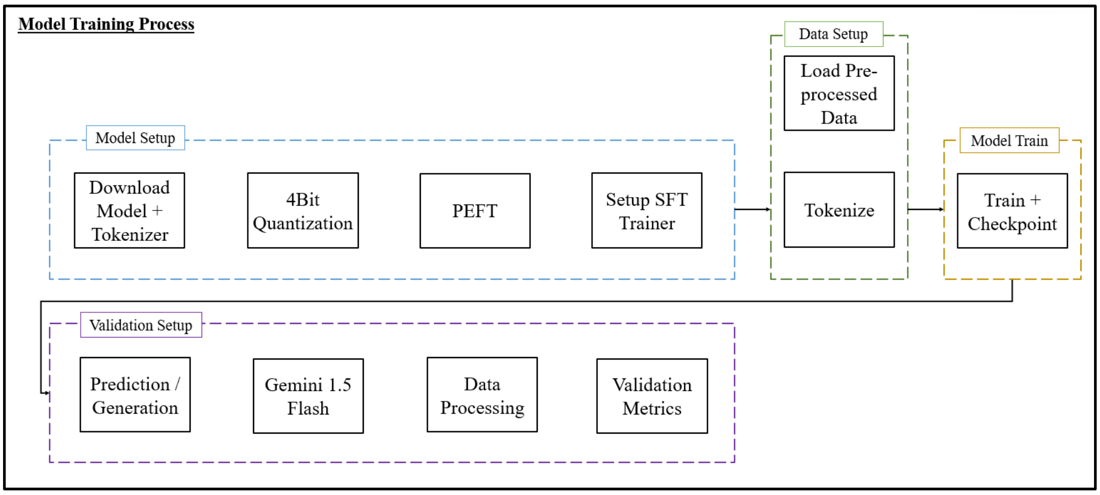
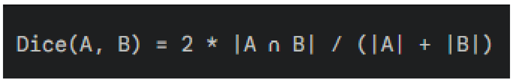


Image 6: Model training process

In the phase 2, data set up, the pre-processed '. jsonl' file for both are loaded and tokenized using the model specific tokenizer.

In phase 3, the model is fine tuned on 10 epochs and default parameters with paged Adam optimizer (J et al., 2024). The dataset is split into 90-10% train-validation sets. For prescription the number of samples for training set were 264 and validation were 30. For seizure frequency the number of samples for training set were 236 and validation set were 27. The reason the validation set is smaller is due to the sparse and small constituency of the dataset that was identified during the EDA. It is possible to perform data stratification and data augmentation technique which is considered in the scope of future work. The model training was conducted on the google collaboratory pro platform. Depending on the GPU selected the training time varied. The faster GPU configuration was expensive. Thus, data augmentation, k fold cross validation and data stratification method were not performed due to time and cost constraints.

In phase 4, in the validation phase, all the check pointed models were reloaded and data were generated for assessment. It was observed that only one of the models generated legible and approximately close output to the expected response. For the other models, the generated output had incomplete JSON dictionary, invalid data dictionary and incomplete generated characters that did not make sense. The approximately closer generated output was then forwarded to Gemini flash 1.5 API, in-order to extract valid JSON. There were multiple JSONs generated by the model. However, only the first valid JSON is extracted. After cleaning the generated output, a valid JSON dictionary was obtained. This dictionary was then used for validation. There were two validation methods implemented to assess the quality and accuracy of the output. The first method of validation, a novelty where the generated dictionary is posed as a multi-label classification problem. There are two components in the dictionary, the key, and its corresponding value. The idea here is to exact match the keys, and the values from the output to the ground truth. The logic for conversion consisted of 3 steps. First step is to create an array of zeroes according to the length of the dictionary. It was found in the EDA that dictionary sizes are inconsistent. Thus, it is important for the array size to be dynamic. Second step is to check if the keys from the ground truth is present in the generated JSON dictionary. If the key is present, then the array is updated with 1 else unchanged as 0. Third step is to perform the step 1-2 on all the dictionary and vertically stack the array to form a matrix and apply selected multi-label classification validation metrics. The selected multi-label classification metrics are exact match ratio, 0/1 loss, hamming loss, precision, recall and F1 score (MMA, 2020). There is no specific criterion for selection of these validation metrics. These were selected as they are generally used metrics in multi-label task. These same steps were applied to obtain an array matrix for values in the dictionary and then validation metrics were applied to it. The second method of validation was to specifically look at the content generated in the 'text' field of the valid JSON. String matching algorithm Sorensen dice coefficient (Dice, 1945) was used to match the text field values. The rationale behind selecting this algorithm with respect to the other string-matching algorithms is because the length of the string is variable.



```
Dice(A, B) = 2 * |A ∩ B| / (|A| + |B|)
```

Image 7: Sorensen Dice Coefficient

The Sorensen dice coefficient considers the length of the string in the denominator (see Image 7). So, in a way, the intersection of the characters from two strings is normalized by the length of the respective input strings. The Sorensen dice coefficient is a value between 0 – 1. The interpretation is higher the coefficient value the better the match between two strings.

RESULTS

It was observed in the result that for the model candidates only one of them was able to generate close to reality JSON structure. The rest of the models generated cluttered responses, random characters, multiple, half and in complete JSON.

Entity	Model		Exact Match Ratio	Hamming Loss	Recall	Precision	F1 Measure
Prescription	Llama 2	Key match	1.00	0	1.00	1.00	1.00
		Value Match	0	0.38	0.62	1.00	0.76
Seizure Frequency	Llama 2	Key match	1.00	0	1.00	1.00	1.00
		Value match	0	0.30	0.60	1.00	0.74

Table 1: Model Result

Table 1 shows the result for Llama 2 model output. From the model result it is observed that the model can correctly generate all the keys of the JSON dictionary. But it struggles to generate the values corresponding to the keys. The comparison is done on model output where it is post-processed to have only one single JSON dictionary.

From Image 8, it could be observed that the content generated in the text field of the JSON, more than 50% of the generated content has less than 0.5 dice score. This is observed in both prescriptions and seizure frequency output. There are certain values where the string generated are exact match with the ground truth. This shows that model is not generating the text specifically indicating the context of seizure frequency or prescription.

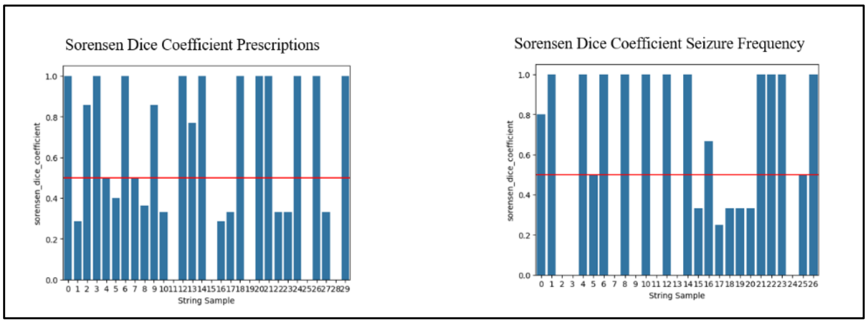


Image 8: Sorensen Dice Coefficient on Validation set

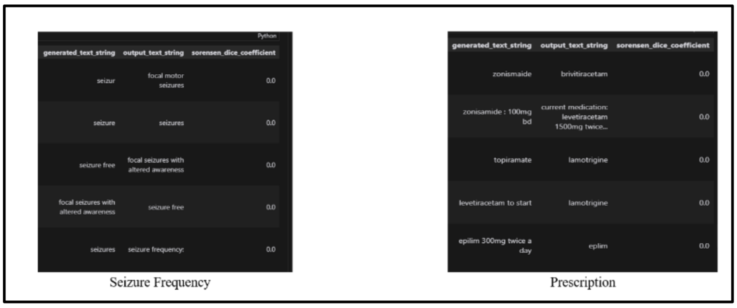
While there are many papers addressing the clinical NER task specific to epilepsy domain (Fonferko-Shadrach et al., 2019; Liu et al., 2020; Xie et al., 2022, 2023) only one of them is directly comparable to the extraction of seizure frequency and prescriptions as both the papers use same dataset. The model performance in comparison to the ExECT (Fonferko-Shadrach et al., 2019) framework is underwhelming. The fine-tuned LLM model underperforms and has lower F1 score in comparison to the rule based and statistical approach developed in the ExECT framework.

DISCUSSION

Initially during the training and fine-tuning stage it was observed that none of the models were generating expected responses. One change that helped to generate expected response at-least in one model was the inclusion of validation steps in the training and fine-tuning phase. It is not ideal that the validation of the model is done on the same set of data that is already exposed to the model. But it became necessary step to experiment and assess the model performance on exposed dataset. The unexpected scenario was the inconsistent and incomplete model output generation for the models except for the Llama 2 model even after exposing the validation set during the training and fine-tuning stage. The other models might work given time for trial-and-error exploration to understand the underlying issues that is resulting in unexpected output. Thus, from this it could be stated that the validation set is not truly a holdout set from the approximate distribution of the training data that could help in true assessment of the model performance.

Since the model output is generated rather than predicted, and due to the involvement of post processing where the generated JSON dictionary is converted to multi-label dummy arrays, it is not possible to get predicted probabilities. Due to non-availability of the predicted probabilities, it is not possible to perform ROC analysis.

Further error analysis was performed to understand the reason behind mismatch between the strings in the text field of the JSON (see Image 9). It was observed that, there were cases where the generated text completely differed than the expected value as seen in the ground truth. There were other cases where the generated text was incomplete and there were cases where the generated text contained the specific word or name of the prescription as found in the ground truth text. This shows that as per the cases, there is a requirement for calibrating the validation method to calculate accurate result. The string-matching algorithm used could be wrapped in the logic that could address the identified cases.



generated_text_string	output_text_string	sorensen_dice_coefficient
seizur	focal motor seizures	0.0
seizure	seizures	0.0
seizure free	focal seizures with altered awareness	0.0
focal seizures with altered awareness	seizure free	0.0
seizures	seizure frequency:	0.0

Seizure Frequency

generated_text_string	output_text_string	sorensen_dice_coefficient
zonismaide	brivitiracetam	0.0
zonisamide : 100mg bd	current medication: levetiracetam 1500mg twice...	0.0
topiramate	lamotrigine	0.0
levetiracetam to start	lamotrigine	0.0
epilim 300mg twice a day	eplim	0.0

Prescription

Image 9: Error analysis on dice coefficients

The future work of the research study could be segregated in three verticals.

First vertical is the dataset. There were problems in the dataset identified during the EDA phase. The dataset was found to be sparse with unequal representation of the sample structures. The dataset was smaller in sample size. Typically, it is preferred to have at least 1000 samples for fine-tuning a LLM model as considered in the industry standard. It is also possible to consider data augmentation technique to increase the sample size of the dataset (Shorten et al., 2021). Feature engineering could be performed to improve the quality of the dataset (Hirose et al., 2024).

The second vertical is the model. The models used for fine-tuning are the medium sized with approximately 7-8 billion parameters. These models are designated as models for industrial and general use. There are another class of models called as research models used for research purposes. These models are bigger in size but has the potential to perform better than the generic class models. Model input followed a chat system-based template, there are other forms of input templates that could be used for designing the prompts as input. The task of NER could be modelled using prompt tuning and prompt engineering approach. These approaches are more cost and time effective approach, but with the drawback of not able to maintain a persistent state of the knowledge or the weights learned during the training phase of the model. Many of the previous literature suggests combination of the methods that mainly include prompt tuning and engineering approach (Brown et al., 2020; Monajatipoor et al., 2024; Munnangi et al., 2024; Naguib et al., 2024). Learning techniques such as zero-shot learning, few-shot learning, in-context learning are the main concepts that will be used for the model to learn from the dataset. To improve the model performance the model could be extended with the frameworks like RAG and its variants (Gao et al., 2024). While the general purpose of the LLMs is to reach a point of artificial general intelligence (AGI) there are small language models (SLM) a decoder-only implementation of transformers architecture could help to achieve machine intelligence that are capable of performing NER like tasks with more cost, time and memory efficient training and fine-tuning (Lu et al., 2024). The finalized Llama 2 model outputs random number of dictionaries. It was observed that, many a times the dictionary was incomplete as well. While in the EDA phase, it was found that there is possibility of more than one JSON annotation in the seizure frequency and prescription cases. So, the current limitation of the model is that it outputs single JSON dictionary, after post-processing. The future work includes to the model output to have more than one JSON dictionary. Since the model generates the JSON keys and corresponding values, the values for the index keys that represent the location of the text values in the clinical letters are also random. It is not understood by the model that the key indicates the location of the generated content in the text field, in the clinical letter. The future work would include to identify the location of the generated text. The selected Llama 2 model was fine-tuned on 10 epochs and default parameters using the paged Adam 8-bit optimizer. Due to cost and time constraints, it was not possible to perform hyperparameter tuning to select the best set of parameters that could include learning rate, weight decay, warm up ratio, gradient accumulation steps, batch size, learning rate scheduler type, different optimizers to find the best result. The future work would include tuning these hyperparameters.

The third vertical is validation. The first method, a novel approach to pose the generated JSON dictionary against ground truth as multi-label classification. There are other methods to explore the accuracy of the generated output. But the validation needs to be performed robustly on true holdout set that approximately represents the distribution of the entire training dataset. The second method, string-matching algorithm was considered on the whole string. Considering the identified cases where the generated text could be a substring of the ground truth string and vice versa or the case where the generated text is a partial match, calibrating the validation logic to consider these cases would give more nuanced validation metric and a perspective to the quality of the generated content in the text field. The generated text could be extended to an interface where the output could be monitored, managed, and annotated by a clinician for rectifying and developing gold standard dataset for future training purposes. However, this is an expensive effort and could be proposed in future work after exploring all other cheaper options.

CONCLUSION

Clinical NER is a challenging task in NLP. The focus on extracting seizure frequency and prescriptions with dosage from the synthetic gold standard epilepsy clinical letters is modelled due to the importance of it in deriving seizure freedom, drug effectiveness and to improve patient care. LLMs a generative model architecture, the latest development in the NLP domain, an implementation and extension of transformer architecture, is used for extracting the entities. The extraction of the entities from the clinical notes into the JSON files is posed as a generation of the JSON files. Data preprocessing is performed to make the dataset ready for model ingestion. The candidate models are accessed and downloaded from hugging face platform. A quantized model coupled with parameter efficient training technique is used for model fine tuning. The dataset is split and limited to generic train and validation sets due to the small size of the dataset. The generated

JSON file is post processed. In the post processing the generated output is sent to Gemini 1.5 flash to clean and extract the valid JSON. Then on the two valid JSON two validation methods are applied. Validation methods include assessing the keys and values of the generated JSON posed as multi-label classification task, and comparing the generated content from the text field of the JSON with the ground truth using the string-matching template. The result suggests that the model underperforms in comparison to the model specifically used for clinical NER in epilepsy and for seizure frequency and prescription extraction task. The future work suggestions indicate the potential of the model to be more scalable and accurate solution than the rule based and statistical implementations. The model could be improved with improvements brought about in the verticals dataset, modelling approaches and customized validation with calibration of the results. The future is only limited to current available research and could be expanded to new methodology and technology as language models are rapidly developing in NLP. The code for the entire research could be found here https://github.com/Xnsam/HDS for open-source, reproducibility, extension of the work, and tracking the current issues until resolution.

REFERENCES

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, *39*(1), 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3

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

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

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

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, *26*(3), 297–302. https://doi.org/10.2307/1932409

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

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

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

Hirose, Y., Uchida, K., & Shirakawa, S. (2024, August 9). *Fine-Tuning LLMs for Automated Feature Engineering*. AutoML Conference 2024 (Workshop Track). https://openreview.net/forum?id=FqbkgaMf8O

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021, October 6). *LoRA: Low-Rank Adaptation of Large Language Models*. International Conference on Learning Representations. https://openreview.net/forum?id=nZeVKeeFYf9

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: JAMIA*, ocad259. https://doi.org/10.1093/jamia/ocad259

J, M. R., VM, K., Warrier, H., & Gupta, Y. (2024). *Fine Tuning LLM for Enterprise: Practical Guidelines and Recommendations* (arXiv:2404.10779). arXiv. http://arxiv.org/abs/2404.10779

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

Jin, Y. (2017). Development of Word Cloud Generator Software Based on Python. *Procedia Engineering*, *174*, 788–792. https://doi.org/10.1016/j.proeng.2017.01.223

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

Kim, J., Lee, J. H., Kim, S., Park, J., & Yoo, K. M. (n.d.). *Memory-Efficient Fine-Tuning of Compressed Large Language Models via sub-4-bit Integer Quantization*.

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

Lu, Z., Li, X., Cai, D., Yi, R., Liu, F., Zhang, X., Lane, N. D., & Xu, M. (2024). *Small Language Models: Survey, Measurements, and Insights* (arXiv:2409.15790). arXiv. http://arxiv.org/abs/2409.15790

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

MMA. (2020, January 25). *Metrics for Multilabel Classification*. Mustafa Murat ARAT. https://mmuratarat.github.io//2020-01-25/multilabel_classification_metrics

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

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

Nasar, Z., Jaffry, S. W., & Malik, M. K. (2022). Named Entity Recognition and Relation Extraction: State-of-the-Art. *ACM Computing Surveys*, *54*(1), 1–39. https://doi.org/10.1145/3445965

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

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

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

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*.

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Text Data Augmentation for Deep Learning. *Journal of Big Data*, *8*(1), 101. https://doi.org/10.1186/s40537-021-00492-0

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

Supervised Fine-tuning Trainer. (n.d.). Retrieved September 26, 2024, from https://huggingface.co/docs/trl/v0.7.4/en/sft_trainer

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. https://doi.org/10.48550/arXiv.2302.13971

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

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

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*.

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

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

"V0.Pdf." Accessed September 29, 2024. https://assets.bwbx.io/documents/users/iqjWHBFdfxIU/r7G7RrtT6rnM/v0.