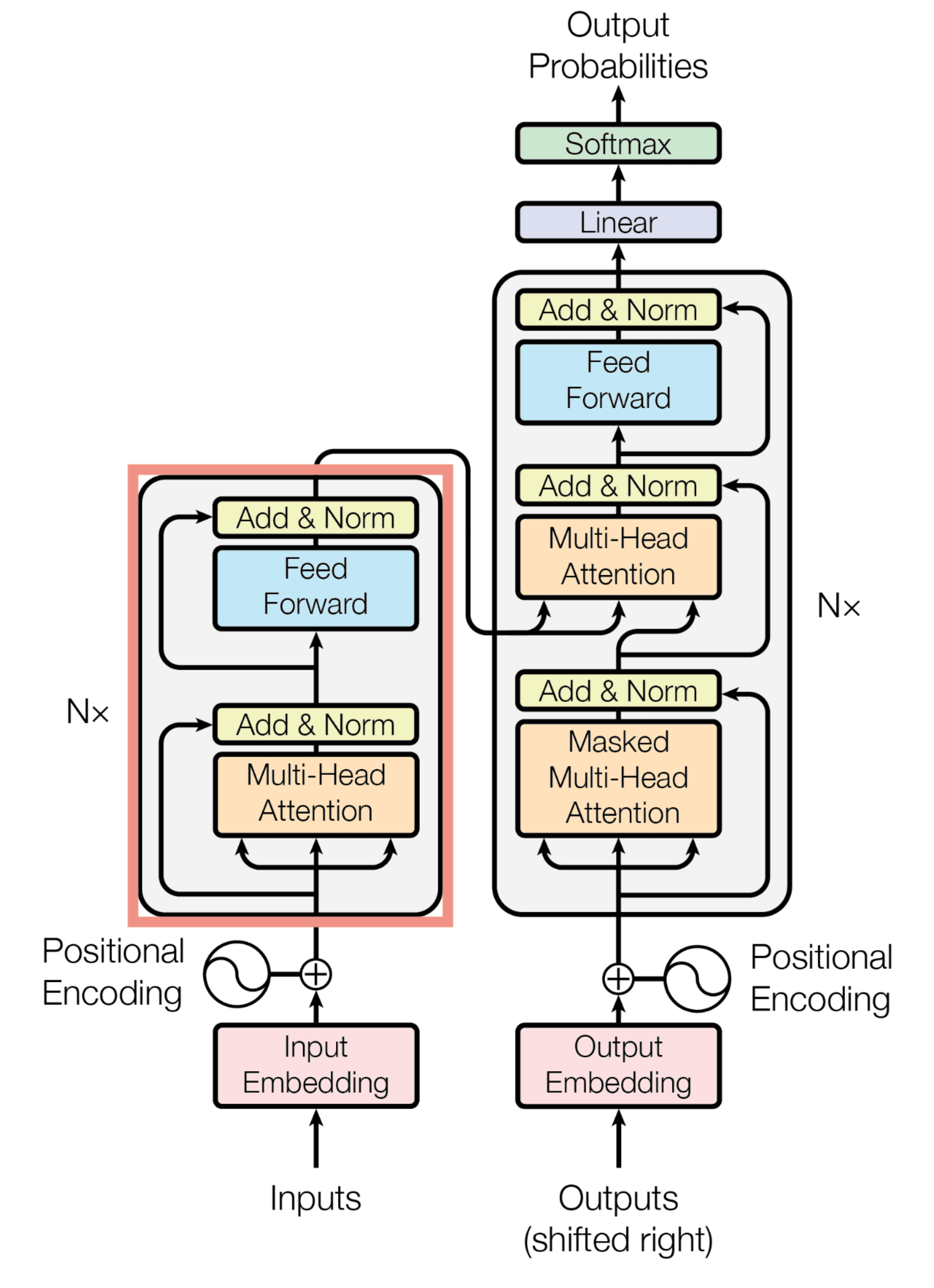



# Assessment of fine-tuned open-source large language models on gold standard synthetic epilepsy clinical notes with focus on Seizure Frequency and Prescriptions

**Presenter**:
Akson Sam Varghese
2311233
M. Sc. Health Data Science
Singleton Park Campus, Swansea University

**GitHub**:
https://github.com/Xnsam/HDS

## Motivation

**Aim**: To perform clinical NER on clinical notes to extract Seizure Frequency (SF) and Prescriptions.

### Why Seizure Frequency ?

1. Seizure Frequency could be used to understand Assess Seizure Freedom and effectiveness of drugs in controlling the seizures
2. Poor documentation and audit trail of seizure frequency robs opportunity to research
3. Capture a rate of seizure to improve clinician-patient engagement and care, study comorbidities e.g. depression induced due to drugs

### Why Prescription ?

1. Research exists in extracting prescriptions from notes for digitization, auditing and automation
2. Crack the existing benchmark with new data extraction approach
3. Nuanced to extract dosage of drugs, to increase or decrease dosage, intake frequency – direct indicators of epilepsy status

References:
Wang et al., 2015; Fonferko-Shadrach et al., 2019; Li et al 2021; Xie et al., 2022; Lehman, E et al., 2023;

## Example

### Given Clinical Text Dataset

Dear Dr,

Diagnosis: symptomatic, structural right temporal lobe epilepsy
Subarachnoid haemorrhage (right MCA) 2017

Current antiepileptic medication: lamotrigine 75 mg twice a day (to increase as stated below)
seizure type and frequency: focal seizures with loss of awareness (Unusual smell) approximately 2 to 3 per month.

Investigations: CT head 2017 collier in situ plus low density right temporal lobe

I reviewed this 57 year old man in clinic today. He continues to have what he calls “absences” which are focal seizures. During them, he has a smell which is difficult to describe. He will lose awareness for a couple of minutes. When he has the warning of an unusual smell he will sit down. He injured his right elbow in a seizure last year.

As he is continuing to have seizures, I suggest that he increases the lamotrigine slowly by 25 mg every fortnight, to an initial dose of 100 mg twice a day. If he has no side-effects with this, then he can again continue to increases the lamotrigine by no more than 25 mg every fortnight.

I will see him again in around six months time, should there be a problem before then he can contact my secretary on the number above.

Prescription

Seizure

### Extract JSON

```
[
  {
   "entity": "Prescription",
   "start_index": "152",
   "end_index": "181",
   "text": "lamotrigine-75-mg-twice-a-day",
   "attributes": {
     "DrugName": "lamotrigine",
     "DrugDose": "75",
     "DoseUnit": "mg",
     "Frequency": "2"
   }
  },
  {
   "entity": "SeizureFrequency",
   "start_index": "239",
   "end_index": "276",
   "text": "focal-seizures-with-loss-of-awareness",
   "attributes": {
     "LowerNumberOfSeizures": "2",
     "UpperNumberOfSeizures": "3",
     "TimePeriod": "Month",
     "NumberOfTimePeriods": "1"
   }
  }
]
```

## **Literature Review** – Search Query Formulation

- Query combination
  - Query combination was formed by using the topics mentioned in the literature search flow chart,
  - Terms parent and child combination along with synonyms
  - Query Structure: Parent term ( AND / OR ) Child Term ( AND / OR )

NLP AND Information Extraction AND Epilepsy AND Clinical NER OR LLMs

Query Examples:

1. Clinical NER with LLMs
2. Information extraction with LLMs
3. Biomedical information extraction with LLMs
4. NLP and clinical NER epilepsy applications
5. Clinical NER with LLMs in epilepsy

# Literature Review – Consult Diagram

**Literature Review** – Results

Overall

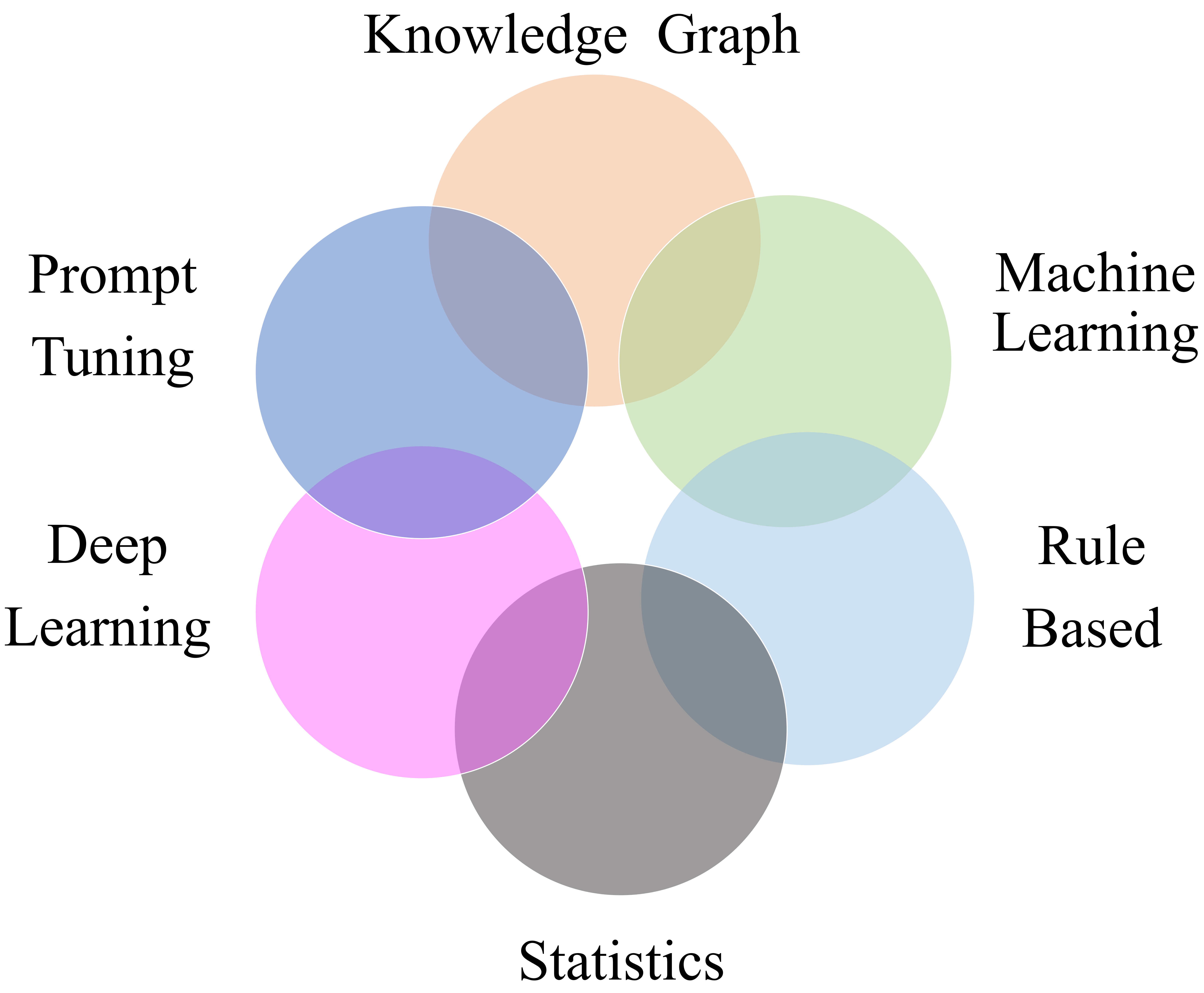
- Total references – NLP: 78; Epilepsy: 19
- Unique Themes – 6
- Combination of Themes – 6+

Pre-LLMs ERA

- Number of papers – 7
- Number of models ( architecture ) overall – 7
- Number of themes – 5

LLMs ERA

- Number of papers – 4
- Number of models ( architecture ) overall – 23
- Number of themes – 3




** See appendix for more detailed metrics on models*

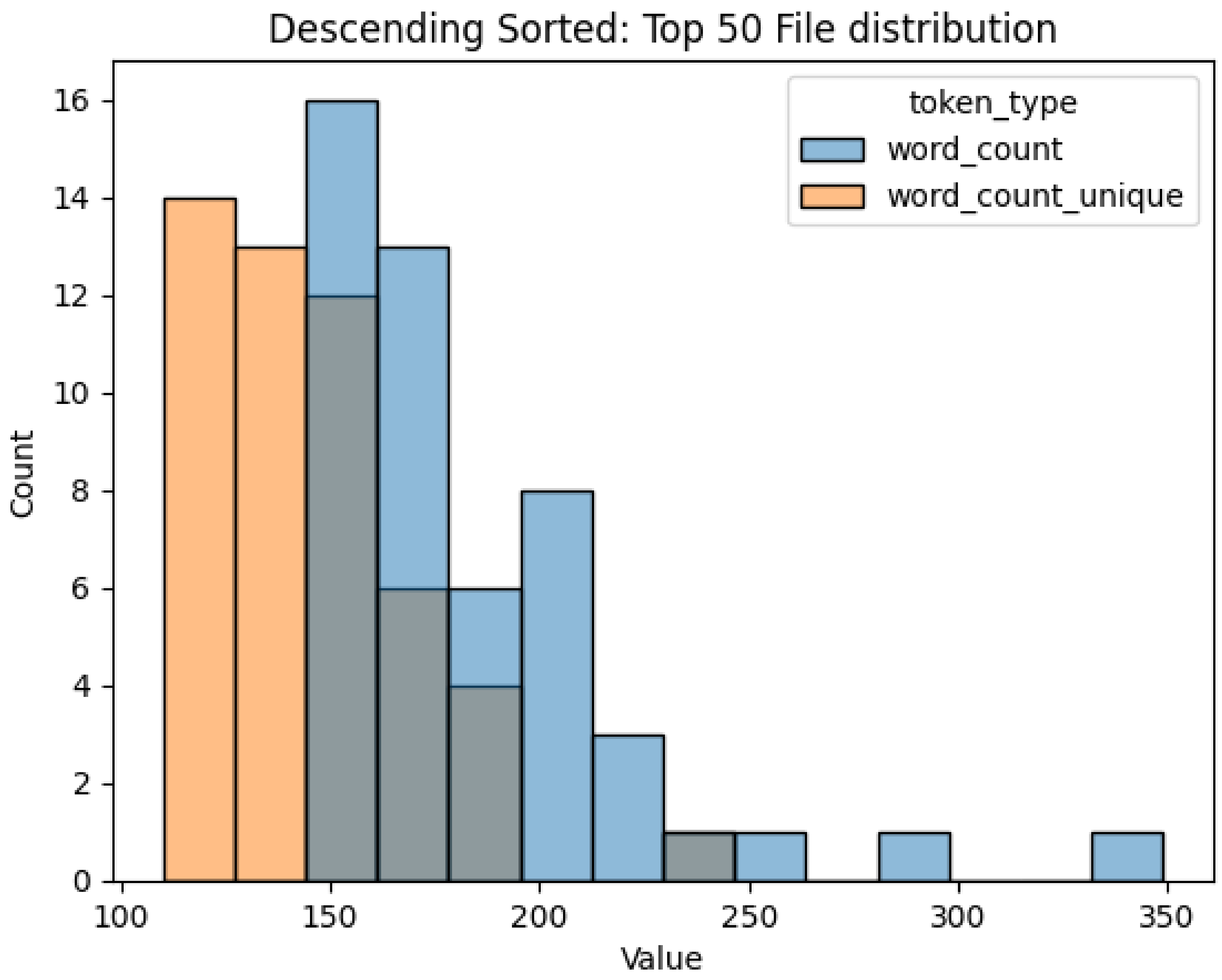
## Development Cycle

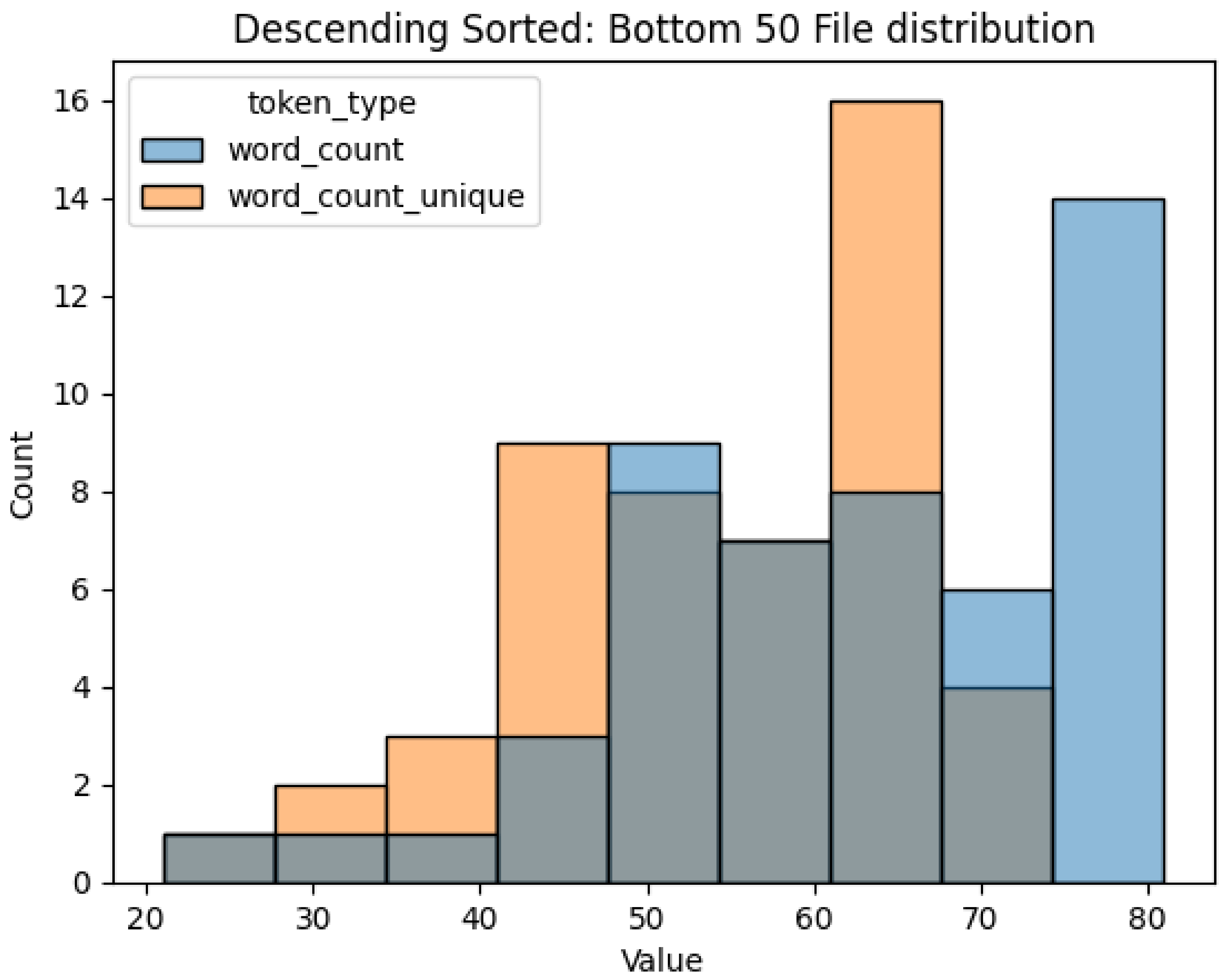

* *Development is an iterative and Semi-cyclic process*

[Gold standard annotation of epilepsy clinic letters for the development of information extraction tools (zenodo.org)]

**Exploratory Data Analysis**

1. Distribution of length of single clinical notes




Insight
1. Identified outlier
2. Average size of the documents with and without unique words in top 50 and bottom 50 files

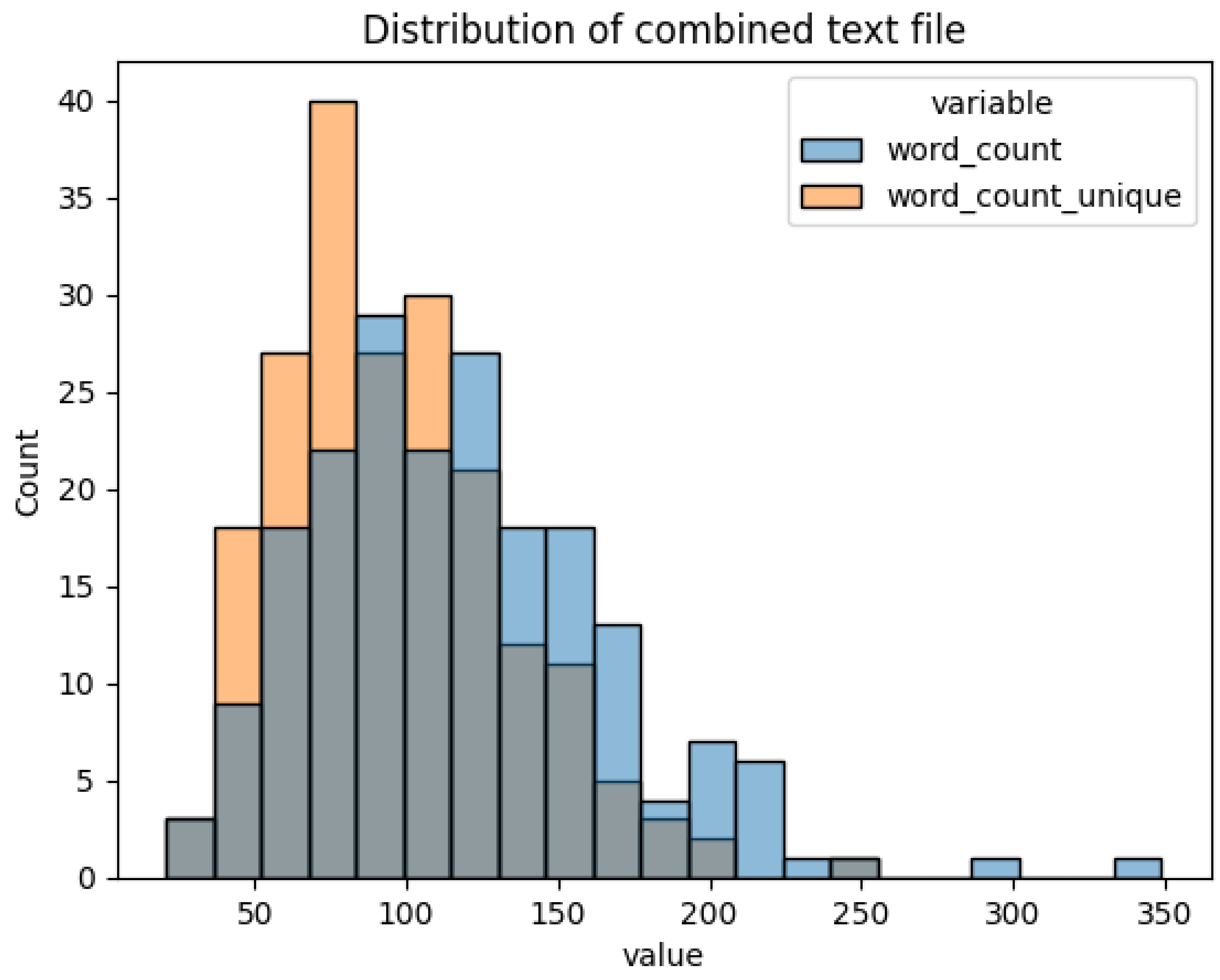
## Exploratory Data Analysis

2. Combined text file: Removed stop words, punctuations, symbols, spaces using standard English and MedSpacy tokenizer, tokenized and counted words

- To understand what should be the max length value in the tokenizer
- If padding / truncation is required in the data

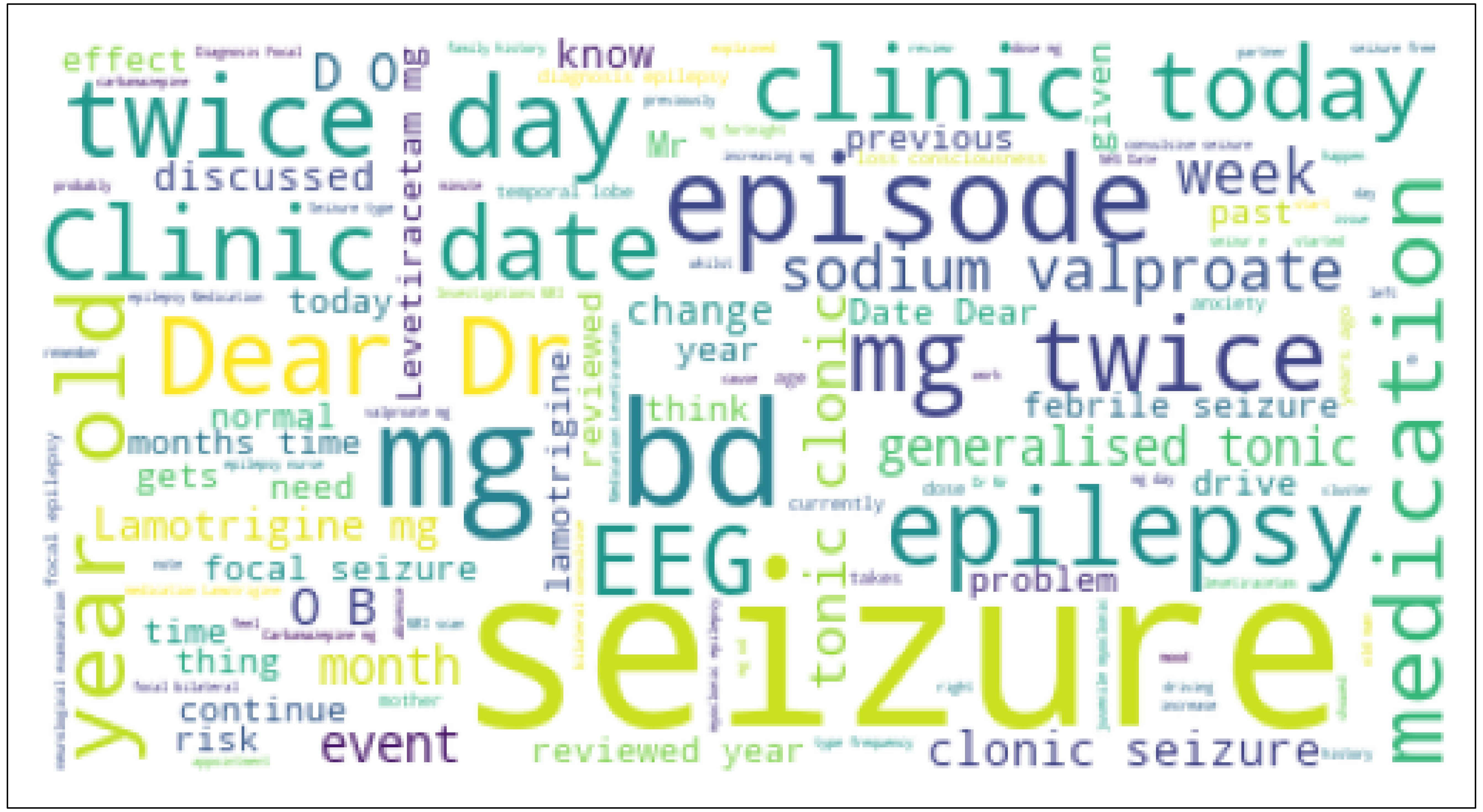
Insight

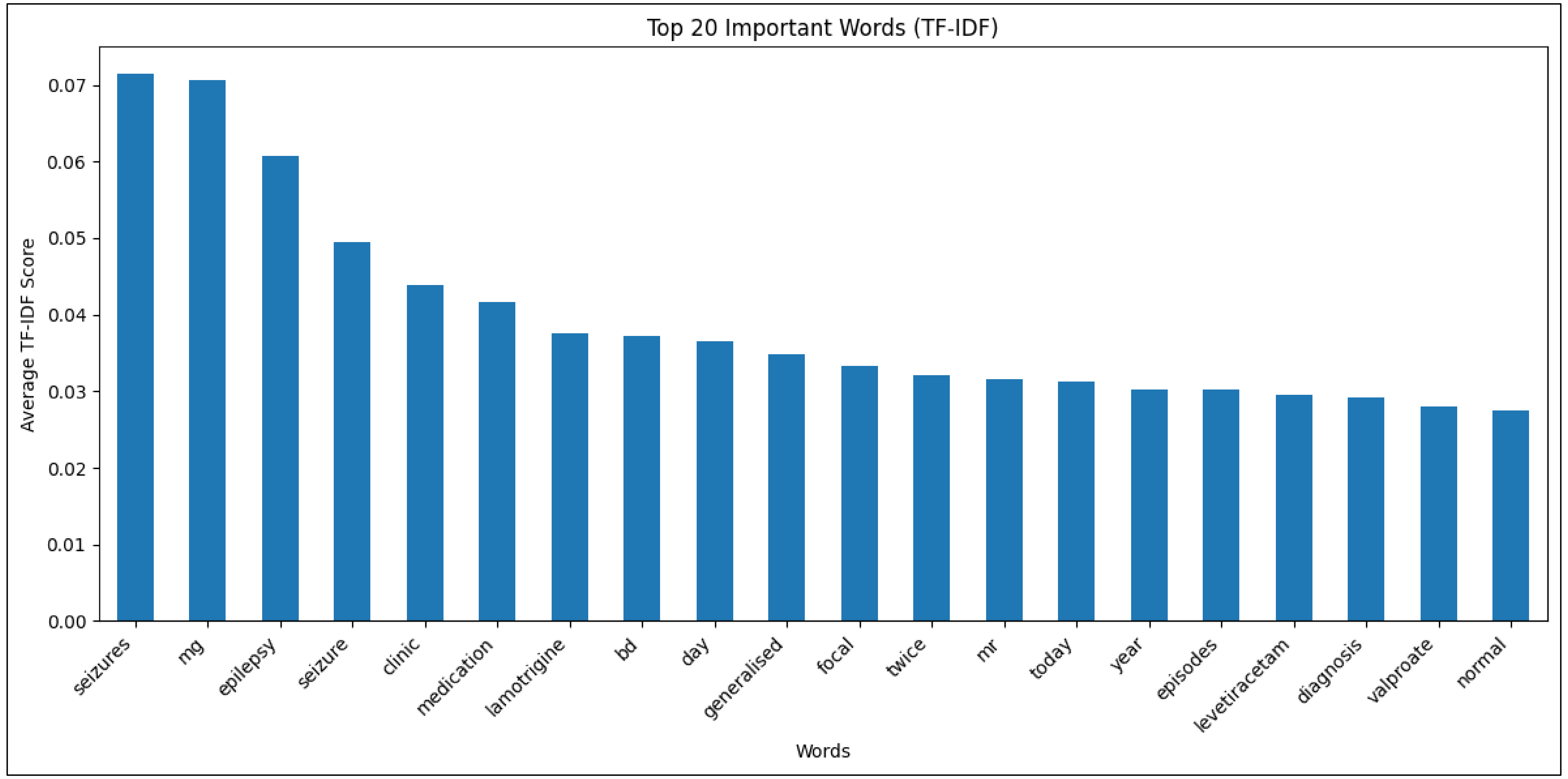
1. Centre and distribution of word in the entire file collection
2. Outlier in the file collection
3. Max length of the string



## **Exploratory Data Analysis**

3. Word Importance




Insight

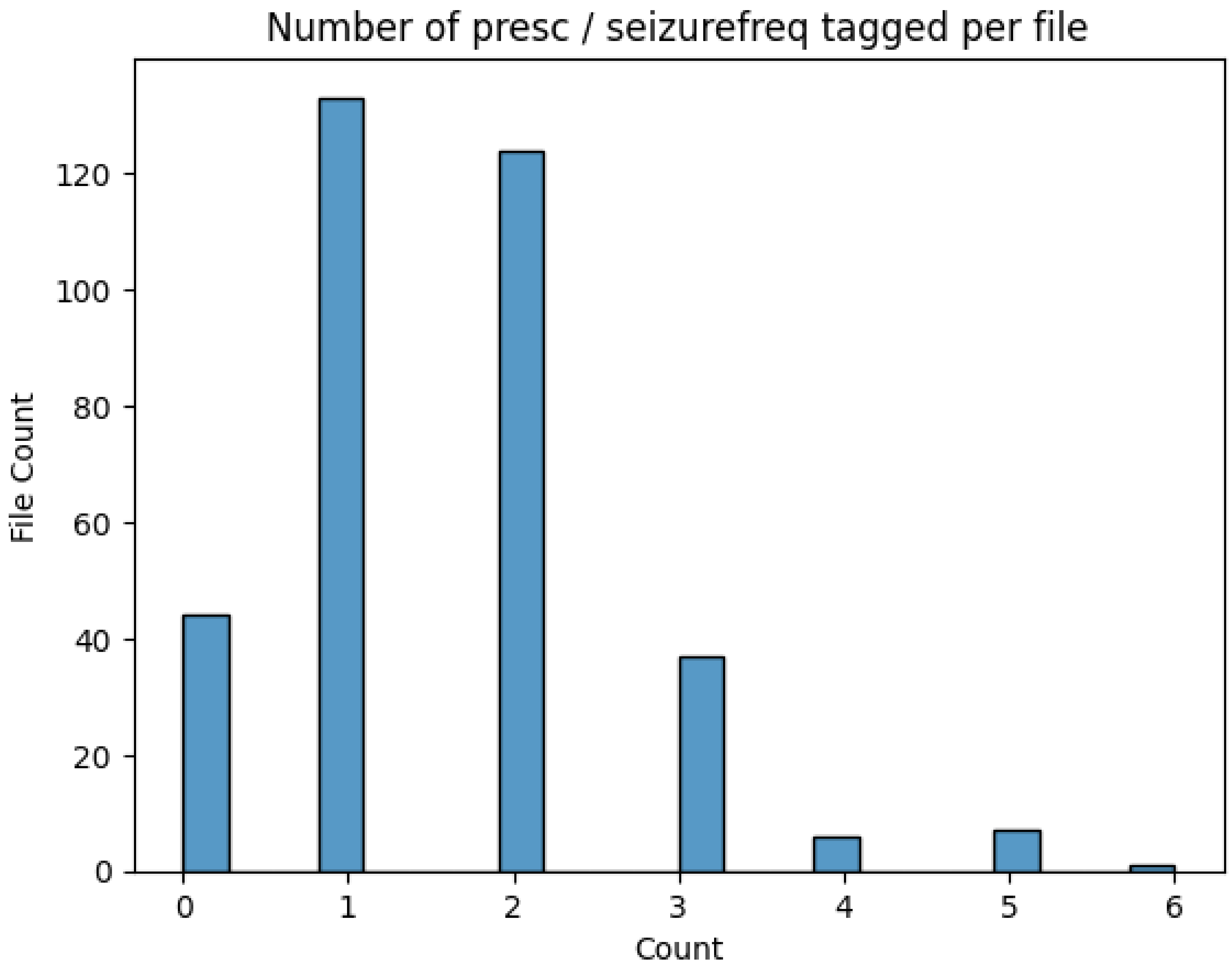
1. Most important word in the entire file collection
2. Top 20 TF – IDF values sorted words in the documents

# Data Preprocess

A. Data Checks
- Prescriptions Absent – 10 / 200
- Seizure Frequency Absent – 34 / 200
- Both Absent – 10 / 200
- Unique JSON structure
  - Seizure – 25 patterns
  - Prescriptions – 3 patterns

Insight
1. Unequal distribution of data in the JSON files
2. Mismatch in data dictionary structure

Action
1. Design separate Input-output pair for the seizure frequency and prescription
2. Create customized structure for each change in the prompt input
3. Data cleaning and simplification is required



## **Data Preprocess**

B. Data Cleaning and simplification


```
[
  {
   "entity": "Diagnosis",
   "start_index": "21",
   "end_index": "73",
   "text": "symptomatic,-structural-right-temporal-lobe-epilepsy",
   "attributes": {...}
  },
  {...},
  {
   "entity": "PatientHistory",
   "start_index": "74",
   "end_index": "98",
   "text": "Subarachnoid-haemorrhage",
   "attributes": {...}
  },
  {
   "entity": "Prescription",
   "start_index": "152",
   "end_index": "181",
   "text": "lamotrigine-75-mg-twice-a-day",
   "attributes": {...}
  },
  {
   "entity": "Diagnosis",
   "start_index": "239",
   "end_index": "276",
   "text": "focal-seizures-with-loss-of-awareness",
   "attributes": {...}
  },
  {
   "entity": "SeizureFrequency",
   "start_index": "239",
   "end_index": "276",
   "text": "focal-seizures-with-loss-of-awareness",
```


1. Filter out only Seizure Frequency and Prescription associated JSON

2. Removed CUI ID

3. Removed the Hyphens from text field in the JSON files

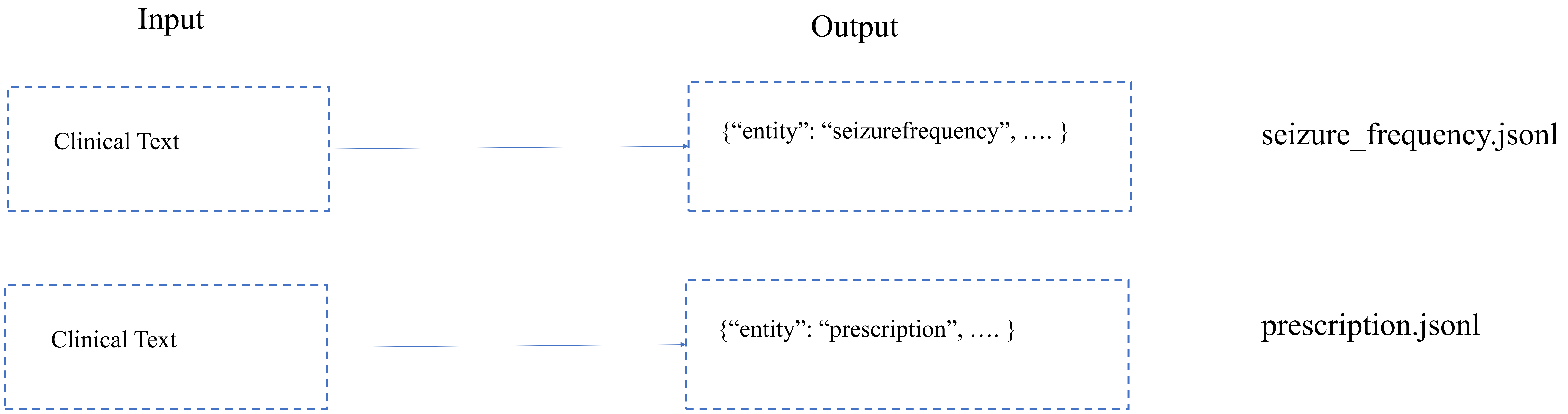
4. Removed “entity” key – value from the JSON

5. Cleaned and Removed Unexpected characters in clinical text
   - Trimmed white spaces and additional tab spaces
   - Removed characters “\u00a0” encoding errors


```
{
  "entity": "Prescription",
  "start_index": "152",
  "end_index": "181",
  "text": "lamotrigine-75-mg-twice-a-day",
  "attributes": {
    "DrugName": "lamotrigine",
    "DrugDose": "75",
    "DoseUnit": "mg",
    "Frequency": "2",
    "CUIPhrase": "lamotrigine",
    "CUI": "C0064636"
  }
},
{
  "entity": "SeizureFrequency",
  "start_index": "239",
  "end_index": "276",
  "text": "focal-seizures-with-loss-of-awareness",
  "attributes": {
    "LowerNumberOfSeizures": "2",
    "UpperNumberOfSeizures": "3",
    "TimePeriod": "Month",
    "NumberOfTimePeriods": "1",
    "CUIPhrase": "focal-seizures-with-loss-of-awareness",
    "CUI": "C0270834"
  }
}
```

## Data Preprocess

C. Input-Output Pairing




JSONL is a text-based format using the . jsonl file extension that is basically the same as JSON format but implemented using newline characters to separate JSON values

**Data Preprocess**

D. Input Templates – Fine Tuning

Instruction Template

~~~
"""Your task is to extract prescription information from a clinical text
Below is the clinical notes from a doctor, delimited by triple quotes.
clinical text: ```{clinical_text}```.
Extract the {extract_entity} only from the clinical text in JSON format.
Give me the output in the json format as mentioned below, delimited by
triple quotes
```{json_format}```.
"""
~~~

Chat System Template

### Human: Your task is to extract prescription information from a clinical text

<<Instruction Template>>

### Answer:

Llama Chat System Input Template

```
[
            {"role": "system", "content": example["instruction"][i]},
            {"role": "assistant", "content": example["output"][i]},
        ]
```

## LLM Model Candidates

Models were built on seizure frequency and prescription separately. Open Source Models are selected.

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

google/gemma-2-2b


Model Training Process
Data Setup
Load Pre-processed Data
Tokenize
Model Setup
Download Model + Tokenizer
4Bit Quantization
PEFT
Setup SFT Trainer
Model Train
Train + Checkpoint
Validation Setup
Prediction / Generation
Gemini 1.5 Flash
Data Processing
Validation Metrics

**Result Validation -** Multi-label classification

- The results were converted to python dictionary and treated as multi-label classification for validation

Dictionary Key Metrics

- Ground Truth keys are the keys of the dictionary and predicted keys are the keys of the generated dictionary

{ “text”: “Levetiracetam”, “start_index”: 1, “end_index”: 20 ….}

Metrics

- Exact match Ratio
- 0/1 loss ( 1 – exact match ratio )
- Hamming Loss
- F1 Score
- Recall
- Precision

Dictionary Value Metrics

- Ground values are the values of the dictionary and predicted values are the values of of the generated dictionary

{ “text”: “Levetiracetam”, “start_index”: 1, “end_index”: 20 ….}

Reference: https://mmuratarat.github.io/2020-01-25/multilabel_classification_metrics

**Result Validation -** Multi-label classification

- Multi-label classification metrics for Prescriptions

| Type | Model | | Exact Match Ratio | Hamming Loss | Recall | Precision | F1 Measure |
|---|---|---|---|---|---|---|---|
| Fine Tuning | Llama 2 | Key match | 1.00 | 0 | 1.00 | 1.00 | 1.00 |
| | | Value Match | 0 | 0.38 | 0.62 | 1.00 | 0.76 |

Size of the train set: 264. Size of the validation set: 30

- Multi-label classification metrics for Seizure Frequency

| Type | Model | | Exact Match Ratio | Hamming Loss | Recall | Precision | F1 Measure |
|---|---|---|---|---|---|---|---|
| Fine Tuning | Llama 2 | Key match | 1.00 | 0 | 1.00 | 1.00 | 1.00 |
| | | Value Match | 0 | 0.30 | 0.60 | 1.00 | 0.74 |

Size of the train set: 236. Size of the validation set: 27

## **Result Validation -** String comparison

- Text comparison between the ground truth and predicted / **generated** text

Ground Truth . : lamotrigine-75-mg-twice-a-day

VS

Generated / Extracted: lamotrigine-75-mg-twice-a-day

### String comparison methods

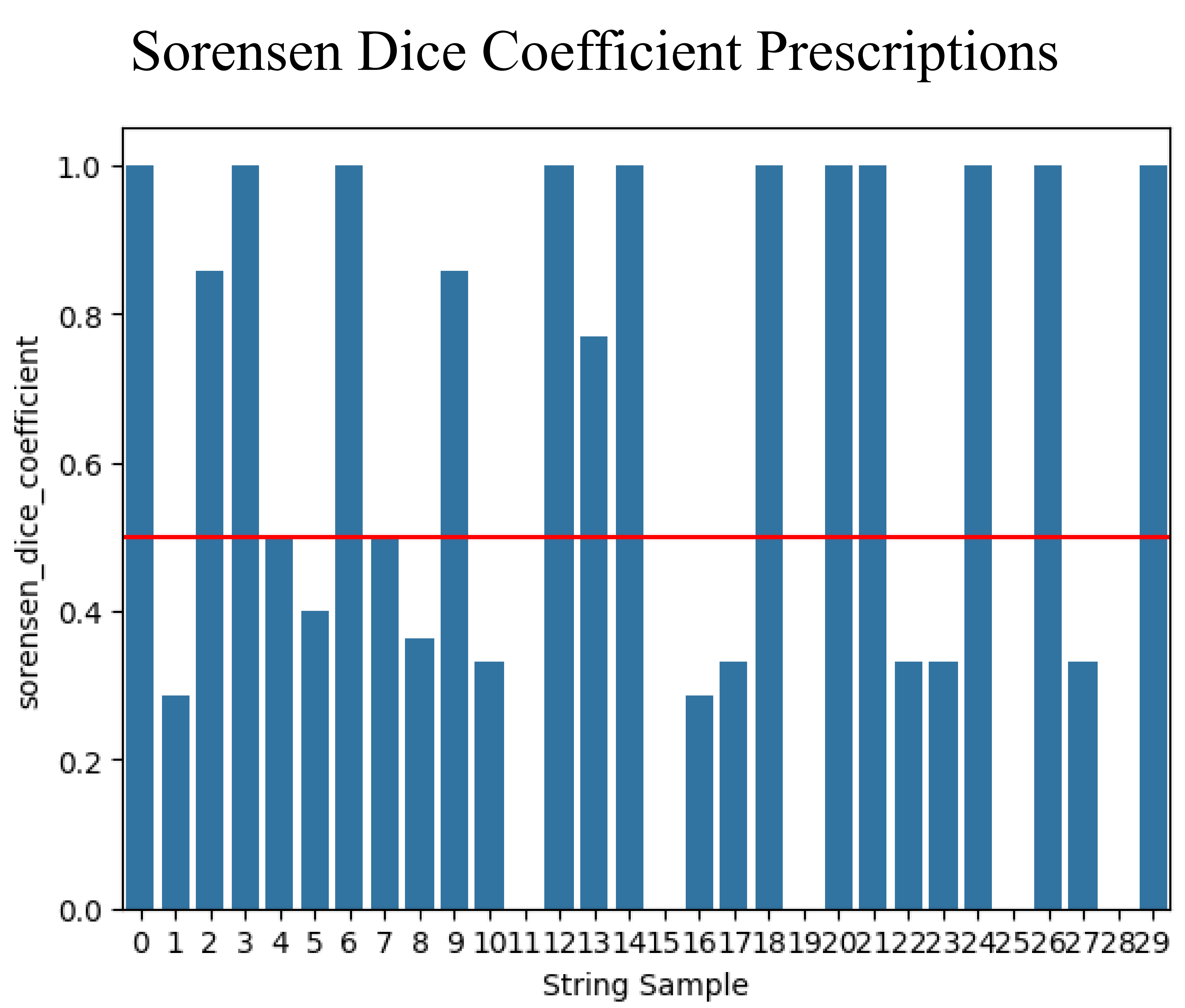
**Sørensen–Dice Coefficient**: A similarity algorithm that computes difference coefficients of adjacent character pairs.

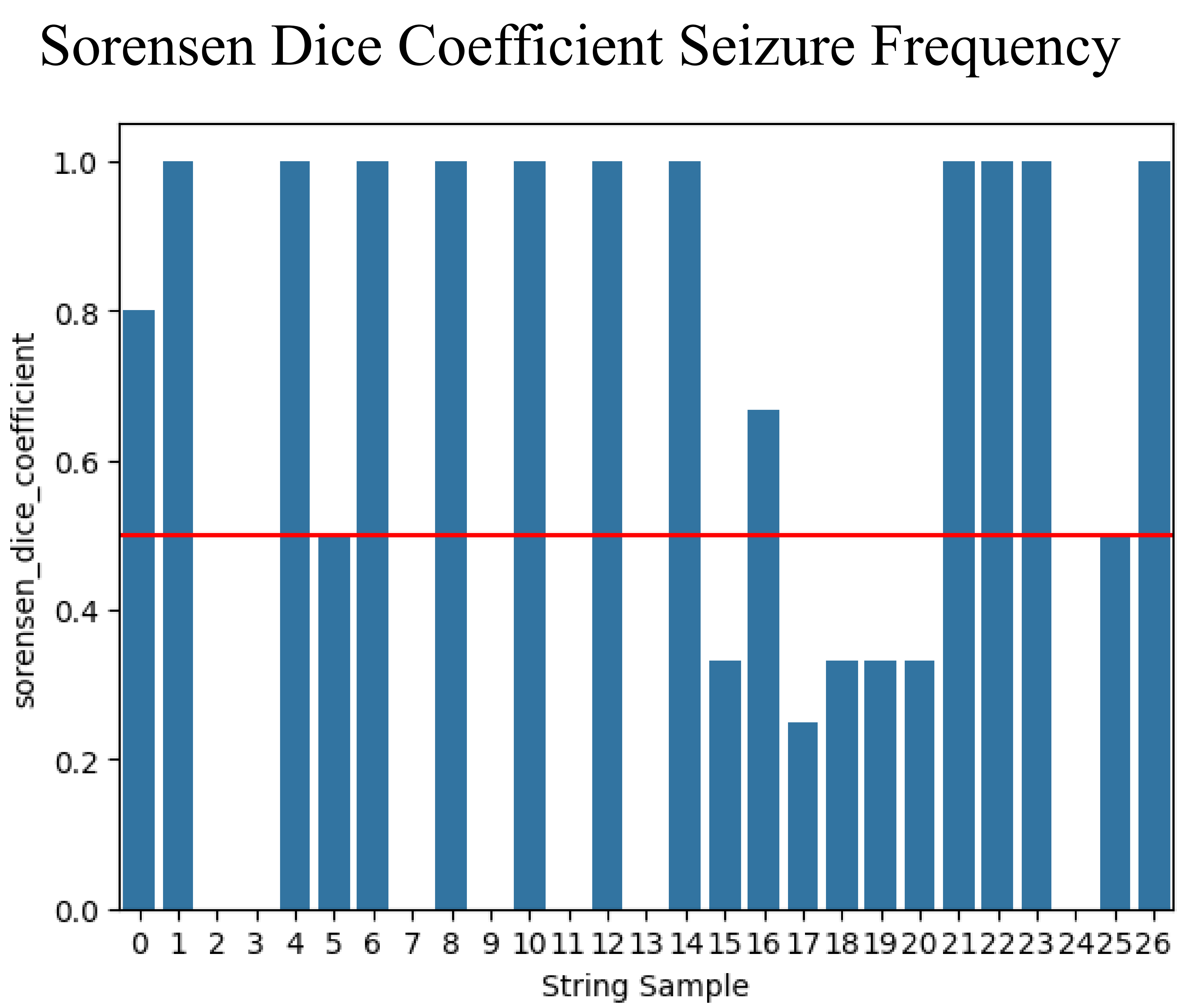
```
Dice(A, B) = 2 * |A ∩ B| / (|A| + |B|)
```

Reference: (Dice, 1945)

## **Result Validation -** String comparison

- Generated text string match – word level




*Threshold: 0.5; 10% validation dataset*

**Error Analysis Sample**

Python

| 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

**Future Work**

### Data

1. Data Augmentation and resampling techniques like random resampling, SMOTE & variants
2. Possibility to increase sample size
3. Feature engineering

### Model

1. Research models – Mixtral 8 x 7B
2. SLM – Microsoft phi3.5
3. Prompt Tuning / Engineering; chain prompting ( COT )*
4. More Compute Time
5. LangChain, RAG & variants approach
6. Output JSON generation - single + multiple
7. Start – End index of text locations
8. Hyperparameter tuning

### Validation

1. Word accuracy calibration
2. Human Intervention for improved accuracy
3. InterpretEval https://arxiv.org/pdf/2011.06854

** Intended to complete by submission*

**Appendix: Literature Review -** Results

Pre-LLMs ERA - Clinical NER tasks

| 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 |

**Appendix: Literature Review** – Results

## LLMs Era - Clinical NER tasks

| 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 |