Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Creating A Gold Standard Annotation Guideline for Electronic Health Records

Joe Davies

joe.m.davies@kcl.ac.uk

King's College London

Beata Fonferko-Shadrach

Swansea University

Ben Holgate

King's College London

Arron Lacey

Swansea University

Huw Strafford

Swansea University

Owen Pickrell

Swansea University

Mark P Richardson

King's College London

Research Article

Keywords: Epilepsy, electronic health records, annotation, inter-annotator agreement, natural language processing

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11693331/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Creating A Gold Standard Annotation Guideline for Electronic Health Records

Joe Davies[1*], Beata Fonferko-Shadrach[2], Ben Holgate[1], Arron Lacey[2], Huw Strafford[2], Owen Pickrell[2], Mark P Richardson[1]

[1*]Basic and Clinical Neuroscience, King's College London, 16 De Crespigny Park, London, SE5 8AF, England.
[2]Faculty of Medicine Health and Life Science, Swansea University, Singleton Park, Swansea, SA2 8PP, Wales.

*Corresponding author(s). E-mail(s): joe.m.davies@kcl.ac.uk;
Contributing authors: B.Fonferko-Shadrach@Swansea.ac.uk; benjamin.holgate@kcl.ac.uk; A.S.Lacey@Swansea.ac.uk; H.A.Strafford@Swansea.ac.uk; W.O.Pickrell@Swansea.ac.uk; mark.richardson@kcl.ac.uk;

Abstract

Purpose: Large Language Models have captured the research environment with their potential to aid in understanding a multitude of natural language processing tasks. Often they require fine-tuning with corpora of relevant information. Electronic health records have an abundance of free-text that is dense with information and could be used for this purpose. This work aims to create a gold-standard annotated dataset with an associated set of guidelines for the annotation process.

Methods: Using a dataset of 1500 electronic health records from epilepsy clinics, we created a set of annotation guidelines and a team of data scientists and clinicians annotated these accordingly.

Results: We find an overall inter-annotator agreement of 0.769, as measured with Krippendorff's Alpha, and showed that the most annotated and agreed upon annotations are those related to anti-seizure medications and diagnoses of epilepsy and other comorbidities.

Conclusion: This work provides a valuable resource for the development and evaluation of NLP systems for epilepsy and electronic health records.

Keywords: Epilepsy, electronic health records, annotation, inter-annotator agreement, natural language processing

1 Introduction

Epilepsy is one of the most common neurological disorders affecting around one in 100 people in the UK and some 51 million world-wide [1]. The disease is characterised by recurring unprovoked seizures of various types, aetiology, and severity. Epilepsy is most commonly treated with anti-seizure medications (ASMs), with the majority of individuals achieving seizure freedom. However, despite the advances made in drug therapies, around 30% of people with epilepsy do not become seizure free. One way to understand why some individuals develop refractory (drug resistant) epilepsy is to study a full, detailed history of the disease together with other life/environmental factors that may contribute to its course, at a population level. The only source of this type of information are the records held by specialist epilepsy/neurology clinics, where patients with epilepsy are investigated, diagnosed, treated, and monitored. They contain details of every patient-specialist encounter including history of seizure onset, seizure type and frequency, ASM dose and side effects, comorbidities, investigation results, and other factors that are reported by patients. Often these records are kept as electronic health records (EHRs) with a large free-text element. These are promising avenues of exploration, then, given their rich wealth of varied information.

Natural Language Processing (NLP) can help understand these EHRs. NLP is the study and development of tools (usually software) that aims to parse human-produced speech. Historically NLP has been used in such fields as: fraud prevention [2][3][4], content moderation [5][6][7], and language translation [8][9][10]. Recently, NLP has also been leveraged to extract information from clinic letters such as: seizure frequency [11], current anti-seizure medications (ASMs) [12], and long-term patterns in seizure freedom [13].

Most of the NLP methods in recent times have been centered around Large Language Models (LLMs). These are deep learning neural networks that are able to encode not just information about how often terms are used, but also the relational information between different terms within a sentence. As a result, they are often able to emulate and parse human-made text with high accuracy. This coupled with the often low barrier for entry in acquiring or querying them means that research can be done quickly and effectively with little start-up cost. Whilst most LLMs are general purpose, there are some that have been specifically tuned for medical purposes such as: MedGemma [14], Meditron [15], and BioMistral [16]. These algorithms score well on common medical tests such as: MedQA [17], MedMCQA [18], and PubMedQA [19]. These are large datasets of questions and answers on topics including epilepsy, cardiology and general medical questions. As a result, these models are often a good starting point for specific research and clinical assistance.

However, when researching techniques using LLMs, it is often found that these models require tuning to be at their most effective [20]. This necessitates the finding of gold-standard, expert-annotated data. There are corpuses of deidentified medical data available, such as: MIMIC-IV [21] (based on patients admitted to intensive care units (ICUs)), NPA-CP [22] (EHRs containing patient information on 11 different pathogens to predict hospital-onset infections), and EHRCon [23] (designed to verify consistency between clinical notes and large-scale relational databases in EHRs). In general, however, there is a lack of freely available EHRs from clinic visits because of constraints

around identifiable patient data.
This project aims to create a gold standard, anonymised, annotated corpus of epilepsy clinical records accompanied by a set of annotation guidelines that can be used by other researchers (in their own analyses). In addition, we will evaluate the inter-annotator agreement (IAA) between those annotating the documents. This acts as a measure of how well the annotation guidelines were followed, and how easy they are to do so. We will also investigate which entities are most and least agreed upon to possibly isolate any problem areas for future investigations.

2 Data

We use 1500 EHRs taken from Guy's and St Thomas' Trust (GSTT). These are EHRs from epilepsy clinics written by clinicians during patient consultations. As a result, the text here is a mix of structured and unstructured. There are consistent layouts for name, address, date of birth etc for the patient, but also sections that describe the interaction, current treatment and future treatment, along with potentially useful insights into the patient's ongoing care.
The letters were selected from a wider corpus using stratified random sampling, focusing on factors such as: age, ASMs present, and seizure frequencies. As a result, we get a dataset that is representative of the center as a whole and, indeed, is a large cross-section of patients with epilepsy (PwE).

3 Method

3.1 Annotation

Data was annotated by 8 users: 5 clinicians and 3 data scientists with epilepsy domain experience. All annotators were employed or associated with King's College London. The annotation tasks were performed using a platform called Markup [22]. Markup provides annotators with a list of entities to annotate, each with their associated attributes. The attributes are either selected from a dropdown menu of options, or inputted by the user. Figure 1 shows an example Markup instance. These entities and attributes are set by a configuration file which is manually loaded onto the system, allowing for bespoke guidelines to be used. All configuration files are made as json-type files. After annotations are completed, the data can be exported in the form of a .ann file, which can be read as an ordinary text file. This is done using the Brat Standoff format [23], ensuring annotations are kept separate from the annotated text. Entities annotated can be found below with a short description of what each entity pertains to:

- Diagnosis: the type of epilepsy, seizures or epilepsy syndrome mentioned
- When Diagnosed: when someone was diagnosed with epilepsy, according to the text
- Onset: age or time since onset
- Epilepsy Aetiology: whether the cause of epilepsy is known or not
- Past History Epilepsy: whether someone has a past history of epilepsy

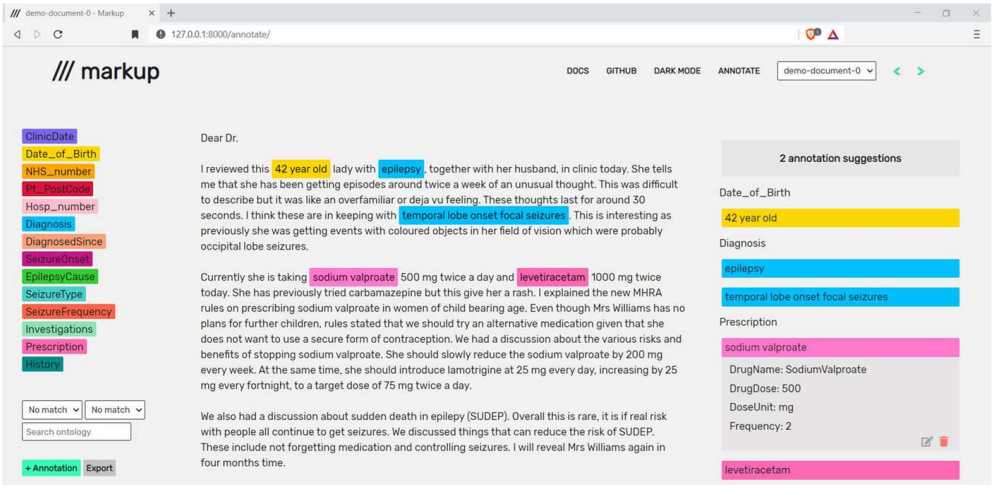


Fig. 1 Example of a letter being annotated in Markup. Text is highlighted and an entity is chosen, with its own list of attributes.

- Seizure Frequency: how many seizures the text suggests someone has. This includes upper and lower limits (eg 2 or 3 per week). Seizure frequency is eventually calculated as seizures per month
- Seizure Free: whether someone is seizure free and how long they have been free for
- Investigations: whether an MRI, CT or EEG has been performed
- ASM: what anti-seizure medications are being taken and information about them (dose, frequency, etc)
- Other Epilepsy Treatments: other treatments such as VNS or ketogenic diets
- ASM Associated Symptoms: any comorbidities directly related to the ASMs being taken
- Past ASMs: any ASMs that have been taken in the past, but that are now not being used for whatever reason
- Other Prescriptions: any non-ASM treatments
- Life Factors: significant life processes that may be affected or affect epilepsy eg pregnancy
- Seizure Triggers: common seizure triggers eg stress, sleep
- Life Style: life style factors that contribute to epilepsy eg smoking, drinking, living on their own
- Family History: whether there is anyone in the family with epilepsy, and the genetic proximity to the individual in question
- Comorbidities: any other diseases present
- Birth History: a normal or abnormal birth and whether they were born prematurely or not

The attributes associated with these entities vary by need, but include both numerical and text inputs. Seizure frequency, for instance, has attributes pertaining to: the date of seizures, how many there were, over how many time periods they occur, how long the

time periods are etc. Diagnostic (pertaining to epilepsy/seizure classifications, comorbidities etc) entities also come with a certainty attribute, allowing a user to attach a value between 1(unlikely) and 3 (definite) inclusive on how likely this entity was what was meant by the original author of the EHR. Terms like "definitely", "probably" and "clearly" would be considered high certainty, whereas "maybe", "possibly" and "somewhat" would indicate low certainty. Entities and attributes were agreed upon in multiple rounds of discussion by clinicians, scientists and patients in both public and private forums.
These entities are annotated item-by-item rather than over the full EHR. This was a decision made such that the process can be considered less mentally taxing on the user, as they just need to annotate as they read rather than annotate after reading through. Instructions for annotation are provided in a guideline document with examples to avoid confusion.
The EHRs were double-annotated in order to provide a method of testing inter-annotator agreement (IAA). The 1500 were split into ten blocks of 150 letters, which were then annotated by one data scientist and one clinician. This was done to test the annotation guidelines and whether they are followable by both those with expertise in clinical work, as well as those that just have domain experience of epilepsy without an explicit background in epilepsy.
All annotations were performed in a trusted research environment.

3.2 Inter-Annotator Agreement

We first group the raw annotations to reduce sparsity, minimise artificial disagreement and calculate clinically relevant IAA. For instance, all information relating to comorbidities — what the comorbidity is, when it began, possible aetiology etc — is grouped into "comorbidity". The annotations come as a mixture of binary (whether an entity is present or not) and non-binary (medication doses, for example) values. Therefore, we then calculate IAA using Krippendorf's Alpha [24]. Krippendorff's Alpha is more suited than other common measures — such as Cohen's Kappa [25] and F1 score [26] — for datasets of both binary and non-binary values. Alpha is defined as:

$$\alpha = 1 - \frac{D_o}{D_e} \tag{1}$$

where: D_o and D_e are observed and expected disagreement respectively. We calculate alpha for each letter between both annotators and, as a result, get the median alpha score for each block of 150 letters. The median was chosen based on the skewness of the distribution of alpha scores, shown in figure 2.

4 Results

The median alpha score across the entire set of entity groups is 0.769, with a range of values being between -0.466 and 1.00. The top 5 raw, ungrouped entities by median alpha score are shown in table 1, and the bottom 5 are shown in table 2. These are both shown with their frequency as a percentage of number of documents annotated in vs total documents.

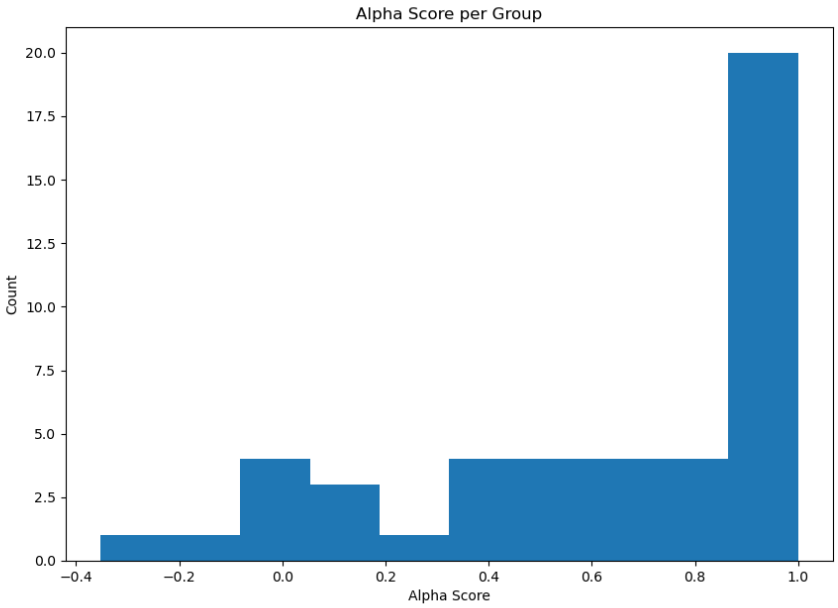


Fig. 2 Distribution of the Krippendorff Alpha scores across all entities.

Entity	Alpha Score	Frequency (%)
Frequency_Levetiracetam	1.00	19.2
DrugDose_Lamotrigine	1.00	13.5
Frequency_Carbamazepine	1.00	8.87
Frequency_SodiumValproate	1.00	8.27
SeizureFree_YearDate	1.00	7.93

Table 1 The top 5 agreed upon raw entities by median alpha score and their associated frequencies. SeizureFrequency_YearDate refers to the year that someone began being seizure free.

Entity	Alpha Score	Frequency (%)
Comorbidities_Certainty	-0.466	6.47
Comorbidities_NumberOfTimePeriods	-0.441	0.281
Frequency_Phenytoin	-0.421	30.0
DoseUnknown_Levetiracetam	-0.001	9.33
DrugDose2_Clonazepam	0.00	0.133

Table 2 The bottom 5 agreed upon raw entities by median alpha score and their associated frequencies. DrugDose2 refers to the second, different dose someone takes as part of their full number per day.

Letters	Median Alpha Score
1-150	0.940
151-300	0.757
301-450	1.00
451-600	0.990
601-750	0.956
751-900	0.495
901-1050	0.520
1051-1200	1.00
1351-1500	0.769

Table 3 Median alpha scores for different spans of letters in 150-large chunks.

Table 3 shows the full median alpha scores for each of the letter blocks.

5 Discussion

The resulting corpus constitutes a reusable annotated dataset for future NLP development. We have shown that annotation guidelines can be followed well to attain good agreement on a carefully sampled set of EHRs. The value for alpha is fairly high at 0.769. Based on accepted guidance, this level of agreement is considered adequate for tentative, but meaningful, conclusions to be drawn from annotated data [27]. However, this does fall below the threshold generally recommended by Krippendorff for high-confidence inference. This is clearly down to outliers in the alpha scores, especially in letter blocks 751-900 and 901-1050. Investigation into these letters shows that they were not particularly different in structure or content than the rest, meaning that any differences likely came from individual raters rather than the records themselves. The annotation guidelines have strict guidance on how to interpret certainty terms, as well as others. The significantly negative value associated with certainty would suggest there is a systematic disagreement between raters. Considering these were the least well annotated, this is likely one cause for low agreement between raters, perhaps if they were not adhering to these terms properly. In addition, one of the main factors of disagreement was around drugs that had multiple different doses e.g. 100 mg in the morning, 150 mg at night. This is something that may well be missed if health records specify a medication is taken twice per day early on, with clarification of different doses happening later. As a result, annotators may miss the specific dosing.
ASM-type annotations being the most numerous and most in agreement is not surprising, considering these are easy to pull out. Oftentimes, prescriptions a patient is using will be listed at the start of an EHR in a pseudo-structured form. Similarly, it is common for other prescriptions, comorbidities and treatments/investigations to be listed early on, adding to their relative ease for annotation.
One factor that was considered was the effect of frequency of annotation vs alpha score. Figure 3 shows a scatter plot of the alpha scores for each grouping annotated against the proportion of documents they are annotated in. We see that there is no clear relationship between the two. In figure 4 we see the distribution of the percentage

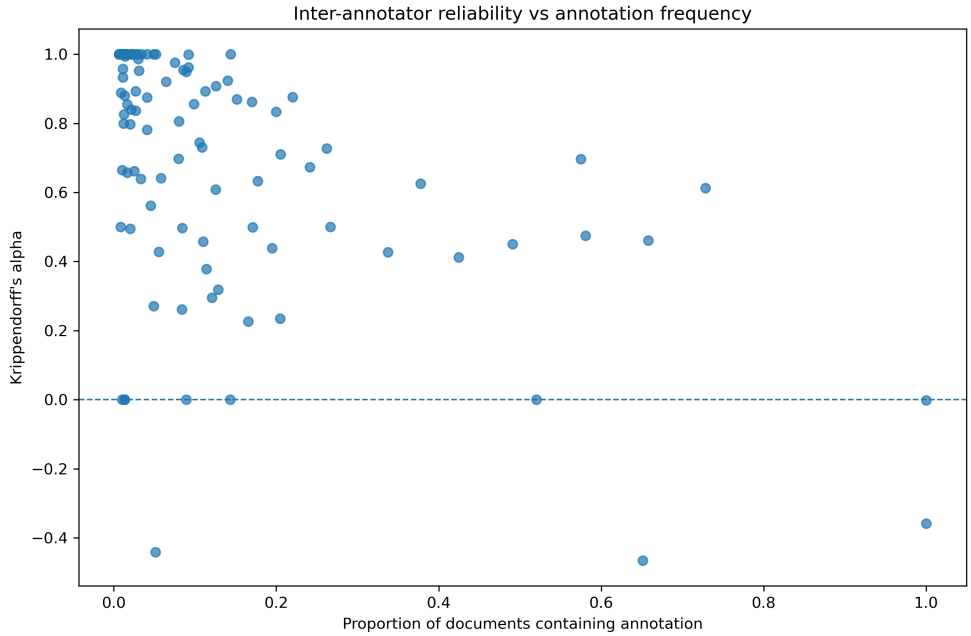


Fig. 3 Alpha score for each group annotated vs the proportion of documents annotated in.

frequency of entities in the best and worst performing groups. There is no significant difference in their shapes, leading us to be able to conclude that the rarity of an entity plays no role in how well it is agreed upon. Overall it would seem that much of the disagreement present is from non-adherence to guidelines and ambiguous entities. Greater effort should be made in the future to impress upon raters that guidelines are to be followed to the letter such that agreements can be more common between raters.

5.1 Future Work

This work is ongoing and other EHRs have already been selected to be annotated. As the corpus grows, we can expect to fine-tune the guidelines and get a greater agreement between raters as a result.
We expect this work to feed into the fine-tuning of LLMs to increase accuracy in information extraction from EHRs in other institutions.

6 Conclusion

This work presents the development of an anonymised, expert-annotated corpus of epilepsy electronic health records together with a set of annotation guidelines designed to support future NLP research. By evaluating IAA across clinician and data scientist annotators, we demonstrate that the guidelines are generally reproducible, achieving a median Krippendorff's alpha of 0.769. While this level of agreement is sufficient for

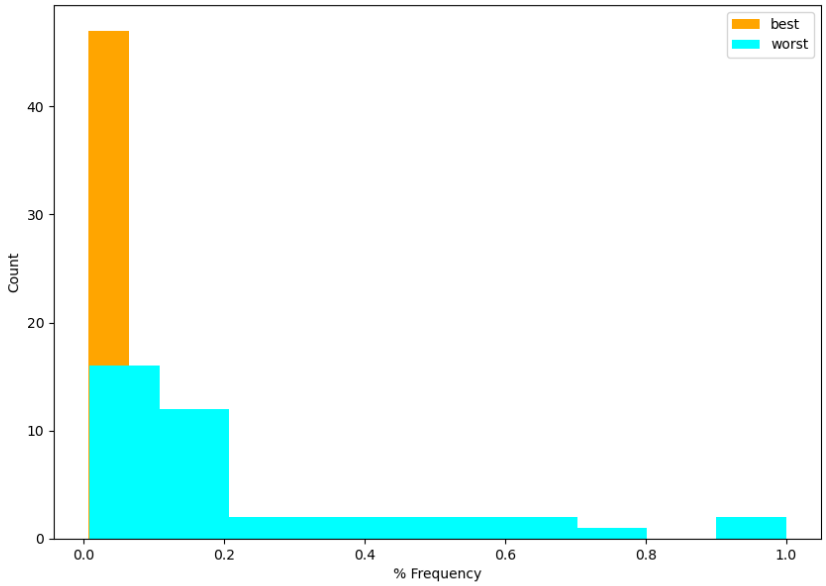


Fig. 4 Distribution of both the best and worst performing entities by percentage frequency of documents annotated.

tentative conclusions, it also highlights opportunities to improve consistency before the corpus is used as a gold-standard resource for model development.

The analysis identified clear patterns in annotation performance. Structured clinical information, particularly ASMs and associated prescribing information, achieved consistently high agreement, indicating that these entities are well defined within both the source documents and the annotation guidelines. In contrast, lower agreement was observed for subjective or ambiguous information, including certainty and temporal expressions. These findings suggest that future iterations of the guidelines should place greater emphasis on handling uncertainty and temporal reasoning, alongside additional annotator training and calibration exercises to reduce variation in interpretation.

Overall, this corpus provides a valuable resource for the development and evaluation of NLP systems for epilepsy EHRs. As annotation continues and the guidelines are refined, the dataset has the potential to support the creation of robust domain-specific language models capable of extracting clinically relevant information at scale. Such tools could facilitate large-scale analyses of epilepsy care and disease progression, ultimately enabling a better understanding of factors associated with treatment response and drug-resistant epilepsy.

Declarations

The authors did not receive any organizational or financial support for the submitted work, including preparation of the manuscript, or conducting the study.
The views expressed are those of the authors and not necessarily those of the National Institute for Health and Care Research, the National Health Service, or the Department of Health and Social Care. All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, the publisher, the editors, or the reviewers.
The authors declare no financial interests/personal relationships which may be considered as potential competing interests.
We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

References

[1] Feigin V et al. Global, regional, and national burden of epilepsy, 1990–2021: a systematic analysis for the Global Burden of Disease Study. The Lancet Public Health. 2025;10:e203–e227.

[2] Bhatt S, Garg G. In: NLP for Fraud Detection and Security in Financial Documents. Springer Nature Switzerland; 2025. p. 131–155.

[3] Chen Y, Zhao C, Xu Y, Nie C, Zhang Y. Deep learning in financial fraud detection: Innovations, challenges, and applications. Data Science and Management. 2026;p. 100162.

[4] Boulieris P, Pavlopoulos J, Xenos A, Vassalos V. Frad detection with natural language processing. Machine Learning. 2023;113:5087–5108.

[5] Mantri, Arjun. Real-Time Content Moderation Using Artificial Intelligence and Machine Learning. International Journal of Science and Research. 2021;10(1):1682–1684. https://doi.org/10.21275/SR24724150350.

[6] Chen C, Qu W, Su S, Feng Y, Li T. A comprehensive review of LLM-based content moderation: advancements, challenges, and future directions. Knowledge-Based Systems. 2025;330:114689.

[7] Heba S, Marton G, Khaled S. Exploring NLP-Based Solutions to Social Media Moderation Challenges. HUman Behavior and Emerging Technologies. 2025;2025:9436490.

[8] L L. Research on the Application of Deep Neural Networks in Machine Translation Method. In: Proceedings of the 2025 International Conference on Machine Learning and Neural Networks. Association for Computing Machinery; 2025. p. 173–177.

[9] Joshi C, Balaji K, Sai B, Abhishek R, Shariff C. Summarization and Translation Using NLP. International Journal for Research in Applied Science & Engineering Technology. 2024;12:555–558.

[10] Team N. Scaling neural machine translation to 200 languages. Nature. 2024;630:841–846.

[11] Holgate B, Davies J, Fang S, Winston J, Teo J, Richardson M P. Fine-tuning LLMs to Extract Epilepsy Seizure Frequency Data from Health Records. In: Proceedings of the 24th Workshop on Biomedical Language Processing. Association for Computational Linguistics; 2025. p. 44–55.

[12] Fang, S, Holgate B, Shek A, Winston J, McWilliam M, Viana P, Teo J, Richardson M P. Extracting epilepsy-related information from unstructured clinic letters using large language models. Epilepsia. 2025;66(9):3369–3384.

[13] Xie K, Gallagher S R, Shinohara R T, Xie S X, Hill C E, Conrad E C, Davis K A, Roth D, Litt B, Ellis C A. Long-term epilepsy outcome dynamics revealed by natural language processing of clinic notes. Epilepsia. 2023;64(7):1900–1909.

[14] Saab K et al.: Capabilities of Gemini Models in Medicine. Https://arxiv.org/abs/2404.18416.

[15] Chen Z.: MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. Https://arxiv.org/abs/2311.16079.

[16] Labrak Y, Bazoge A, Morin E, Gourraud P-A, Rouvier M, Dufour R.: BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. Https://arxiv.org/abs/2402.10373.

[17] Jin D, Pan E, Oufattole N, Weng W-H, Fang H, Szolovits P.: What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. Available from: https://arxiv.org/abs/2009.13081.

[18] Pal A, Umapathi L K, Sankarasubbu M. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In: Proceedings of the Conference on Health, Inference, and Learning. vol. 174 of Proceedings of Machine Learning Research. PMLR; 2022. p. 248–260. Available from: https://proceedings.mlr.press/v174/pal22a.html.

[19] Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. PubMedQA: A Dataset for Biomedical Research Question Answering. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics; 2019. p. 2567–2577.

[20] Anisuzzaman D, Malins J, Friedman P, Attia Z. Fine-Tuning Large Language Models for Specialized Use Cases. Mayo Clinic Proceedings: Digital Health. 2025;3(1):100184.

[21] Kraljevic Z et al. Validating Transformers for Redaction of Text from Electronic Health Records in Real-World Healthcare. In: 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI); 2023. p. 544–549.

[22] Fonferko-Shadrach F et al. Annotation of epilepsy clinic letters for natural language processing. Journal of Biomedical Semantics. 2024;15.

[23] Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. brat: a Web-based Tool for NLP-Assisted Text Annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics; 2012. p. 102–107.

[24] K K. Content Analysis: An Introduction to Its Methodology. SAGE Publications, Inc.; 2019.

[25] Cohen J. A Coefficient of agreement for nominal Scales. Educational and Psychological Measurement. 1960;20:37–46.

[26] C R. Information Retrieval. Butterworths; 1979.

[27] Marzi G MD Balzano M. K-Alpha Calculator–Krippendorff's Alpha Calculator: A user-friendly tool for computing Krippendorff's Alpha inter-rater reliability coefficient. MethodsX. 2024;12:102545. https://doi.org/https://doi.org/10.1016/j.mex.2023.102545.