A Framework for Flexible Extraction of Clinical Event Contextual Properties from Electronic Health Records

Shubham Agarwal[1], Tom Searle[1], Mart Ratas[1], Anthony Shek[2], James Teo[2,3], Richard Dobson[1,4]

[1]King's College London, London, UK
[2]Guy's and St Thomas' NHS Foundation Trust, London, UK
[3]King's College Hospital NHS Foundation Trust, London, UK
[4]Health Data Research UK and University College London, London, UK

Correspondence: shubham.agarwal@kcl.ac.uk

Abstract

Electronic Health Records contain vast amounts of valuable clinical data, much of which is stored as unstructured text. Extracting meaningful clinical events (e.g., disorders, symptoms, findings, medications, and procedures etc.) in *context* within real-world healthcare settings is crucial for enabling downstream applications such as disease prediction, clinical coding for billing and decision support. After Named Entity Recognition and Linking (NER+L) methodology, the identified concepts need to be further classified (i.e. contextualized) for distinct properties such as their relevance to the patient, their temporal and negated status for meaningful clinical use. We present a solution that, using an existing NER+L approach - MedCAT, classifies and contextualizes medical entities at scale. We evaluate the NLP approaches through 14 distinct real-world clinical text classification projects, testing our suite of models tailored to different clinical NLP needs. For tasks requiring high minority class recall, BERT proves the most effective when coupled with class imbalance mitigation techniques, outperforming Bi-LSTM with up to 28%. For majority class focused tasks, Bi-LSTM offers a lightweight alternative with, on average, 32% faster training time and lower computational cost. Importantly, these tools are integrated into an openly available library, enabling users to select the best model for their specific downstream applications.

1 Introduction

Electronic Health Records (EHRs) document patient interactions, health data, and treatment details, including secondary uses for non-clinical, administrative, or research purposes (NHS, 2023). This data is stored in various formats, with unstructured text comprising a significant portion (Häyrinen et al., 2008). Clinical text classification is a vital step in the sequence of tasks that facilitate the extraction of clinical information. These tasks can unlock tremendous opportunities for large-scale systemic analysis (Spasic et al., 2020), ranging from the detection and prediction of adverse events (Tayefi et al., 2021), to the coding of cancer pathology reports (Tayefi et al., 2021) and improving the quality of care (Menachemi and Collum, 2011), among numerous others.

Before text classification, we perform a Named Entity Recognition and Linking task (NER+L) to extract clinical events such as a diagnosis, symptom, finding or procedure, and link each span to a standardised clinical terminology. For example, in the text "patient has been confirmed a diagnosis of diabetes", the NER+L task will extract the entity 'diabetes' as the diagnosis 'diabetes mellitus' and link, for example, the SNOMED CT (SNOMED) identifier: SCTID: 73211009.

For this, we build on the existing MedCAT (Kraljevic et al., 2021) implementation which is part of the CogStack (Jackson et al., 2018) ecosystem. MedCAT is an openly available and easily fine-tunable NER+L tool designed for large-scale clinical text processing which is integrated within the CogStack framework, a scalable platform for processing unstructured EHR data in real-world healthcare environments. Appendix A.5 outlines the Cogstack ecosystem and the MedCAT frameworks for training and inference.

After NER+L, further contextualization is required to ensure that the extracted entities capture the context in which the entity appears. This can be referred to as an entity attribute (Savova et al., 2010), property, modifier or a meta-annotation in the MedCAT context. The modifier categories we consider in this work are:

- Presence: (Not present | Hypothetical | Present) - to determine if the entity is negated, positively or hypothetically mentioned.
- Experiencer: (Other | Family | Patient) - to

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 946–959

determine if the entity was experienced by the patient, family member or is referred to in some other way.

- Temporality: (Past | Future | Recent) - to determine the time of the entity

The above tasks provide essential contextual information whilst being suitably flexible for a range of downstream uses. The most frequent use is to filter only those clinical events that are Presence: *Present*, Experiencer: *Patient* and Temporality: *Recent*. Figure 1 describes an example clinical text and the modifier classification output.
In the context of MedCAT, this contextualization task is referred to as MetaCAT.

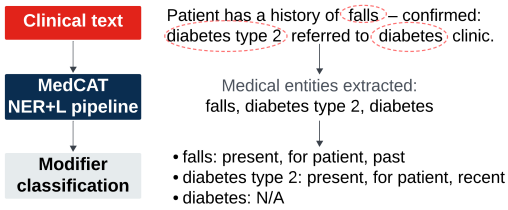

Figure 1: Example output for context modifier classification

Text classification, particularly in the medical domain, is challenging due to the complexity of the data, the extensive use of medical jargon, the sensitive nature of the information, and the presence of inconsistent or missing data (Ratwani, 2017). Additionally, medical data often suffers from class imbalance, presenting further challenges (Khushi et al., 2021).
To address these challenges, prior work has explored the use of Bi-directional Long-Short Term Memory (Bi-LSTM) (Mascio et al., 2020), transformer approaches, i.e Bidirectional Encoder Representations (BERT) (Devlin et al., 2019) models (Li et al., 2024) (Si et al., 2019) and causal large language models (Nazi and Peng, 2024).

In this study, we analyze and present a deployed NLP solution within an the CogStack-MedCAT framework for large-scale classification and contextualization of medical entities across a diverse range of clinical NER+L projects. This ensures that extracted entities are accurately categorized within their clinical context, improving reliability for downstream tasks. Specifically, we:

- Evaluate the performance of Bi-LSTM, Masked language models (BERT, ModernBERT) and larger Causal language models (Llama, Mistral) for clinical text classification on real-world EHR data.

- Analyze the impact of class imbalance and explore mitigation techniques to enhance performance for underrepresented classes.

- Leverage Large Language Models (LLMs) to generate synthetic data and investigate in-context learning for medical classification tasks.

- Provide comprehensive tooling to users to train, evaluate and use trained models for specific and often varied downstream uses.

Our work contributes to the deployment of NLP in healthcare by addressing practical challenges such as scalability, adaptability, and model performance in real-world clinical settings where extracted clinical events are often mixed and diverse, and tools are deployed and used in often low compute availability settings.

Table 1: Dataset description

Category	Class	Samples
Presence	Not present (False)	578
	Hypothetical (N/A)	978
	Present (True)	7430
Experiencer	Other	1002
	Family	75
	Patient	7908
Temporality	Past	733
	Future	484
	Recent	7771

2 Methodology

2.1 Dataset Description

The dataset is sourced from CogStack, deployed at Guy's & St Thomas' NHS Foundation Trust and comprises of 14 annotation projects, 1800 documents, 10252 annotations, and 203 distinct clinical events across the 3 tasks.

The data has been collected across multiple clinical specialties and clinical operational use cases e.g. geriatrics, nephrology, ENT and metabolic disorders. Table 1 shows the aggregate distribution of annotations across all projects.

2.2 Masked Language Models

In this study, we use a BERT (Devlin et al., 2019) model, a Transformer (Vaswani et al., 2017) based encoder only model as our base model to perform the described medical text classification task (bert-base-uncased)[1]. From early experimentation, incorporating the representation of the entire sequence along with the medical entity improved performance over just including the embedding representing of the medical entity. We used the BERT model with 10 encoder layers, trained with a dropout rate of 0.2, the AdamW optimizer combined with a learning rate scheduler, and a batch size of 128. Stratified splitting is employed to all trained models to ensure that all classes are adequately represented in both the training and test datasets.

In this study, we experiment with frozen BERT parameters and fine tuning BERT with LoRA (Liu et al., 2022). Our experiments show LoRA-based fine-tuning enables effective model adaptation. This model configuration is ablated with alternative methodologies described in Section 2.4.

In addition to BERT, we evaluate ModernBERT as well, given its improvements over standard BERT in general-domain NLP tasks (Warner et al., 2024). This allows us to assess whether recent improvements translate to medical text classification.

2.3 Bi-LSTM Model

We also employ a Bi-LSTM model for the given classification task. In this workflow, the text inputs are tokenized using Byte-Level Byte-Pair Encoding (BBPE), a subword-level tokenizer adapted for word segmentation (Sennrich et al., 2015; Wang et al., 2020a; Wolf et al., 2019). The resulting tokens were embedded using pretrained Word2Vec (Mikolov et al., 2013) embeddings, which were fine-tuned during training to better suit the task-specific vocabulary and semantics. Training was conducted using the AdamW optimizer, with a dropout rate of 0.3, 5 Bi-LSTM layers, and a batch size of 128.

2.4 Class imbalance

Class imbalance is a common challenge in real-world datasets, particularly in clinical data (Kumar et al., 2022). Our dataset exemplifies this, as for the Experiencer task, the 'Family' class represents only 1% of the data compared to the 'Patient' class. Despite efforts to collect additional annotated data for underrepresented classes, the class distribution remained unchanged, highlighting the issue of class imbalance. To address class imbalance, we use the below mentioned methodologies with the masked language models and the Bi-LSTM model.

2.4.1 Class Weights

Class weights can address class imbalance by giving different weights (importance) to the majority and minority classes. The difference in class weights impacts training by assigning higher weights to the minority class to penalize its misclassification while reducing the weight for the majority class encourages the model to learn and better recognize the minority classes (Johnson and Khoshgoftaar, 2019).

2.4.2 Synthetic Data Generation using LLM

One potential solution to class imbalance is to generate additional data for the underrepresented classes. We use the Mistral 7B instruct model (Jiang et al., 2023) for data generation as in our experimentation, it demonstrated superior data generation capabilities compared to Llama 3 (Dubey et al., 2024). The model is prompted with 10 examples from our manually collected dataset, 8 from the minority classes and 2 from the majority classes. Manual validation was performed to ensure the integrity of the data. The synthetic data comprises less than 5% of the total dataset, which prevents the data distribution from being significantly altered. We randomly sample clinical events to generate synthetic examples for each of the 3 tasks. Appendix A.1 shows examples of generated data for all tasks.

2.4.3 2-Phase Learning

2-phase learning (Lee et al., 2016) is a training approach designed to fix the issue of the gradients being dominated by the majority class. Each phase varies class weights usage and learning rate resulting in majority class dominance being mitigated. The 2 phases in this approach are:

- Phase 1: In this phase, all classes are down sampled to a specified value N (that is close to the number of samples for the minority class) and training is performed with higher class weights given to minority classes. Phase 1 allows the model to capture and learn the details for the minority classes.

[1] https://huggingface.co/google-bert/bert-base-uncased

- Phase 2: During this stage, the model undergoes a second round of training, now on the entire dataset. The class weights assigned to minority classes are high but lower compared to the initial phase. This phase allows the model to capture the finer details for all classes, leading to a more finely-tuned model.

2.5 Causal Large Language Models for classification

Causal Large Language Models (LLMs) have seen widespread usage in NLP and specifically in text classification tasks (Spasic et al., 2020). We use Llama 3.1 8B instruct (Dubey et al., 2024) and Mistral 7B instruct (Jiang et al., 2023). These models have been pre-trained on large volumes of web-scale data (Brown et al., 2020), then further pre-trained to follow instructions (Brown et al., 2020).

For classification, we rely solely on **zero-shot** and **few-shot learning**, as the high computational cost makes large-scale fine-tuning infeasible at our clinical sites where compute resources are limited. Zero-shot learning (Radford et al., 2019) (Larochelle et al., 2008) is where the model performs classification based only on the instructions in the prompt without any 'training' examples (Rohrbach et al., 2011). In few-shot learning (Wang et al., 2020b), the model is prompted with a limited set of examples (inputs and their corresponding outputs) alongside the classification instructions, enabling it to better understand the task at hand. For few-shot learning, the models were provided with a total of 9 examples, distributed as 3 examples per class. The choice of 9 examples per task aims to maintain simplicity, clarity, and conciseness in the prompts, with longer prompts having the potential to reduce the model's effectiveness in performing these tasks (Brown et al., 2020) (Sahoo et al., 2024). Appendix A.2 contains the prompts used for both models. For practical use in real-world applications, we consider the trade-offs of using LLMs, including model size, performance and computational resource requirements.

3 Results

This section reports model performance using macro F1-score and recall, which are particularly relevant given the severe class imbalance. Table 2 summarizes the results for all tasks, while Appendix A.4 presents the ablation results for each task.

3.1 Performance of Models

BERT models consistently achieved higher macro F1-score and minority class recall compared to both Bi-LSTM and ModernBERT models.
Bi-LSTM models, when combined with class imbalance mitigation techniques, showed improved performance for one minority class but struggled on the other. In contrast, BERT models demonstrated consistently strong performance across both minority classes, achieving up to 28% higher recall for minority classes.

ModernBERT also benefited from class imbalance mitigation and performed well across both minority and majority classes. However, BERT model achieves higher macro F1-score and recall for minority class on all classification tasks. This performance gap can be attributed to ModernBERT's design optimizations for efficiency, which could limit its capacity to capture the complex contextual relationships often present in medical text.

3.2 Performance of Class Imbalance Mitigation Techniques

Synthetic data generation consistently improved minority class recall, especially in the Experiencer and Presence tasks. However, this did not translate into an improved macro F1-score and in many cases reduced performance on majority class.

2-phase learning led to enhancements in both BiLSTM and BERT models for F1-score and especially recall for minority classes, which improved up to 9%. In most cases, it outperformed synthetic data generation, suggesting it is more effective at addressing class imbalance.

The combined approach of synthetic data and two-phase learning outperformed all other setups across models and tasks. In addition to improving minority class recall, it also boosted macro F1-score and majority class performance in several cases, indicating a more balanced and generalizable learning process. Notably, it achieved gains with up to 16% improvement in minority class recall and 11% improvement in macro F1-score for the Experience task.

3.3 Performance of LLMs for in-context classification

This section evaluates the performance of Llama and Mistral models in few-shot learning for our classification tasks. As zero-shot learning produced subpar results, we plan to report on en-

CW - class weights in favour of minority classes; 2PL - 2-phase learning fine-tuning approach + CW; SD - inclusion of synthetically generated data + CW
* indicates the majority class for the task.
w/ = with

Table 2: Model performance for all classification tasks

Task	Model	Accuracy	Macro F1-score	Recall		
				Not present	***N/A***	***Present****
Presence	Bi-LSTM (w/ 2PL + SD)	**0.89**	0.84	0.84	0.79	0.92
	BERT (w/ 2PL + SD)	**0.89**	**0.87**	**0.87**	**0.84**	0.9
	ModernBERT (w/ 2PL + SD)	**0.89**	0.85	0.86	0.8	0.93
	Llama 3.1 8B (few shot)	0.84	0.45	0.6	0.03	**0.97**
	Mistral 7B (few shot)	0.8	0.38	0.1	0.2	0.95
				Other	***Family***	***Patient****
Experiencer	Bi-LSTM (w/ 2PL + SD)	0.92	0.83	0.84	0.73	0.93
	BERT (w/ 2PL + SD)	**0.93**	**0.93**	**0.89**	**0.94**	**0.95**
	ModernBERT (w/ 2PL + SD)	**0.93**	0.87	0.83	0.84	**0.95**
	Llama 3.1 8B (few shot)	0.69	0.51	0.05	0.9	0.75
	Mistral 7B (few shot)	0.74	0.53	0.17	0.65	0.8
				Past	***Future***	***Recent****
Temporality	Bi-LSTM (w/ 2PL + SD)	**0.91**	0.84	0.75	0.84	0.93
	BERT (w/ CW)	0.82	0.8	0.8	0.78	0.83
	BERT (w/ 2PL + SD)	0.87	**0.86**	**0.84**	**0.86**	0.89
	ModernBERT (w/ CW)	0.86	0.8	0.7	0.81	0.91
	ModernBERT (w/ 2PL + SD)	0.92	0.84	0.79	**0.86**	**0.94**
	Llama 3.1 8B (few shot)	0.8	0.43	0.1	0.36	0.9
	Mistral 7B (few shot)	0.77	0.47	0.27	0.55	0.74

hanced performance after applying the techniques discussed in Section 4.5. Both Llama and Mistral models showed performance limitations, particularly for minority classes, as indicated by their low macro F1-scores and recall. The lowest recall value observed was 0.05 for the Experiencer category (achieved by Llama). However, both models performed well on the majority class, with Llama reaching a high recall value of 0.97 for the Presence task. While few-shot offers advantages, it did not yield optimal results. Further analysis is performed in Section 4.2.

4 Discussion

4.1 Class Imbalance Mitigation Techniques

Our analysis highlights the varying strengths of the three imbalance mitigation strategies tested. Synthetic data generation enhanced minority class performance by increasing training exposure for these classes. However, its impact was limited as models frequently misclassified minority instances as majority class labels. This highlights the need for complementary strategies as synthetic data generation alone is insufficient to overcome strong learning biases.

2-phase learning first trained models on a balanced subset to ensure early exposure to all classes, helping them prioritize minority class patterns before majority class dominated training. While this led to improved performance for recall and macro F1-score, its impact was limited by the small size, narrow coverage and low diversity of minority class examples in the balanced subset, reducing the model's ability to generalize to more complex

instances.

The combined approach of synthetic data generation and 2-phase learning yielded the strongest performance on recall and macro F1-score across all tasks and models. This combination works effectively because the techniques complement each other well: synthetic data generation enriches the representation of minority classes, ensuring the model is exposed to sufficient and varied examples; and 2-phase learning then allows the model to focus on minority classes first - now with a richer and more diverse set of examples, enhancing performance on these before fine-tuning on the full dataset. This combined approach ensures balanced performance making the model more effective and reliable in real-world healthcare text classification tasks.

4.2 LLMs for in-context classification

4.2.1 Performance for classification

LLMs for in-context classification exhibited limitations in consistently classifying minority classes with high recall, except for specific cases (e.g., Llama for the Experiencer task). Our investigation revealed that this is likely due to a bias towards the majority class, where the LLMs tend to classify a sample as the default class (majority class) unless there are clear and explicit indicators of the minority class. This approach struggles as the indicators for minority classes are often subtle and contextual, not always explicit. In healthcare settings, where nuanced language is common, this bias poses challenges for accurately classifying clinical events.

4.2.2 Deployment challenges

Deploying LLMs in real-world applications poses challenges, primarily due to their high computational cost. While fine-tuning LLMs would allow for a fairer comparison with other methods, it is largely impractical given the substantial compute and time requirements involved. Hence in-context learning is considered due to its ability to be used directly for inference.

Although in-context learning with LLMs eliminates the need for labeled data and excels in majority class performance, these benefits are outweighed by model size, inference cost, and real-time deployment challenges. From experience, our typical project will assess multiple years of EHR data, potentially looking to classify many tens of thousands of clinical events for their contextual attributes. More widely running these models over the entirety of multi-decade EHR records will involve millions of potential contextual classifications, which is challenging in healthcare IT settings due to hardware constraints.

4.3 Classification Task Analysis

The modifier classification tasks are essential for contextualizing medical entities, ensuring accurate presence, attribution, and timing, which enhances clinical decision support by reducing misinformation. We analyzed these classification tasks to understand the complexity each task poses in real-world healthcare settings. The models performed best on the Experiencer task due to clear class boundaries. The Presence task was more challenging, as 'Not present' and 'N/A' can overlap despite conceptual differences. The Temporality task was the most difficult, with 'Recent' being well-defined, while 'Future' and 'Past' varied widely in time range and often lacked explicit quantification, adding to the complexity.

4.4 Beyond Experimentation: Real-world applications in Healthcare

4.4.1 Typical workflow for NLP project

Typically, clinical academic researchers or a healthcare data analyst will present a research question or project. This will first define a set of relevant EHR data. The project will then evaluate, fine-tune and run provided models to extract a structured and contextualized representation of the unstructured clinical data.

4.4.2 Real-world deployment projects

This system is deployed in multiple healthcare projects, including: early detection of high-risk Chronic Kidney Disease patients, identification of Brugada Syndrome cases, and the Fluoropyrimidine Audit, where majority-class performance is critical. These projects leverage structured entity classification to enhance risk stratification, patient outcomes, and clinical workflows. We have numerous projects where the minority classes of specific tasks provide an important distinction. E.g. the 'Armed Forces Identification' looks to identify relatives of military personnel (Experiencer: *Family*), 'Cardio Myopathy' aims to identify prognosis (Temporality: *Future*).

The findings of this study provide guidance for real-world deployment: for projects where majority class performance is the primary focus—such as the Fluoropyrimidine Audit, Bi-LSTM presents a

viable choice due to its lower computational cost, faster training time and high performance on majority class. Conversely, BERT is the most reliable option when identifying minority classes is critical, as it consistently outperforms Bi-LSTM and ModernBERT in recall for underrepresented categories. BERT's higher computation cost and higher training time (up to 32% slower) is justified by its superior overall and specifically minority class performance, while Bi-LSTM offers a lightweight solution for majority class tasks. ModernBERT is more efficient than BERT but sacrifices some ability to capture complex medical contexts. For tasks requiring high accuracy, especially with minority classes, BERT remains the better choice.

These insights enable the development of a suite of models tailored to different needs and use cases, supporting scalable, high-accuracy NLP applications with significant implications for patient care.

4.5 Limitations and Future Work

The data for this study was sourced from a single, albeit multi-hospital provider site. We plan to expand our dataset and run further experiments across multiple sites, supporting more diverse use cases of these models. We used the '*bert-base*' variant in this study. We will incorporate '*bert-large*' and domain-specific models such as ClinicalBERT (Huang et al., 2019) and BioBERT (Lee et al., 2020) as they can improve performance. We also plan further experiments with ModernBERT to explore potential improvements and evaluate its performance with all class imbalance mitigation techniques across tasks. For in-context classification with LLMs, we plan to: tweak the prompts to encourage the inclusion of subtle indicators of minority classes, investigate the impact of using higher number of samples per class for few-shot prompting on performance and also utilize Human-in-the-loop and Chain-of-thought prompting techniques to boost performance (Wei et al., 2022). Furthermore, we intend to explore the parameter-efficient approach of prompt tuning (Lester et al., 2021), which enables task adaptation without fine-tuning the model. This method is well-suited to settings with limited computational resources and provides a more practical and equitable comparison with the fine-tuning approaches discussed.

5 Conclusion

The BERT model, combined with synthetic data generation using LLMs and 2-phase learning, delivered the best performance, particularly in improving recall for minority classes. This highlights an effective strategy for addressing class imbalance in medical text classification. This research contributes to the field of medical NLP by developing a suite of models tailored to diverse use cases for extracting clinical event data from unstructured medical text, thereby enhancing clinical decision support and patient care.

Acknowledgments

We appreciate the help and support from GSTT, the GSTT-Cogstack team and Aleksandra Foy for helping with data collection in this work. This work was supported by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. SA, TS, RD are part-funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. RD is also supported by The National Institute for Health Research University College London Hospitals Biomedical Research Centre. This paper represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Ethics

We handle EHR data with strict governance protocols, adhering to institutional and legal guidelines. Due to the sensitivity of the data, it cannot be freely shared for replication, and our datasets are available only on-premise to minimize the risk of data leakage and unauthorized access. Given the real-world impact, we recognize the risks of algorithmic bias and misclassification, especially for minority classes, and mitigate this through class imbalance techniques and thorough evaluations. However, our models are intended to support, not replace,

clinical expertise. By integrating our tools into an open-source framework, we support accessibility, reproducibility, and ethical AI deployment in healthcare.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kristiina Häyrinen, Kaija Saranto, and Pirkko Nykänen. 2008. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International journal of medical informatics*, 77(5):291–304.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Richard Jackson, Ismail Kartoglu, Clive Stringer, Genevieve Gorrell, Angus Roberts, Xingyi Song, Honghan Wu, Asha Agrawal, Kenneth Lui, Tudor Groza, and 1 others. 2018. Cogstack-experiences of deploying integrated information retrieval and extraction services in a large national health service foundation trust hospital. *BMC medical informatics and decision making*, 18:1–13.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54.

Matloob Khushi, Kamran Shaukat, Talha Mahboob Alam, Ibrahim A Hameed, Shahadat Uddin, Suhuai Luo, Xiaoyan Yang, and Maranatha Consuelo Reyes. 2021. A comparative performance analysis of data resampling methods on imbalance medical data. *IEEE Access*, 9:109960–109975.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, and 1 others. 2021. Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117:102083.

Vinod Kumar, Gotam Singh Lalotra, Ponnusamy Sasikala, Dharmendra Singh Rajput, Rajesh Kaluri, Kuruva Lakshmanna, Mohammad Shorfuzzaman, Abdulmajeed Alsufyani, and Mueen Uddin. 2022. Addressing binary classification over class imbalanced clinical datasets using computationally intelligent techniques. In *Healthcare*, volume 10, page 1293. MDPI.

Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. 2008. Zero-data learning of new tasks. In *AAAI*, volume 1, page 3.

Hansang Lee, Minseok Park, and Junmo Kim. 2016. Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning. In *2016 IEEE international conference on image processing (ICIP)*, pages 3713–3717. IEEE.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Yiming Li, Wei Tao, Zehan Li, Zenan Sun, Fang Li, Susan Fenton, Hua Xu, and Cui Tao. 2024. Artificial intelligence-powered pharmacovigilance: A review of machine and deep learning in clinical text-based adverse drug event detection for benchmark datasets. *Journal of Biomedical Informatics*, page 104621.

Haoran Liu, Ziyi Qin, Nian Xu, Yuxuan Wu, Zhiheng Bao, Shengqi Guo, Hang Peng, Jian Chen, and Jun Zhou. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Aurelie Mascio, Zeljko Kraljevic, Daniel Bean, Richard Dobson, Robert Stewart, Rebecca Bendayan, and Angus Roberts. 2020. Comparative analysis of text classification approaches in electronic health records. *arXiv preprint arXiv:2005.06624*.

Nir Menachemi and Taleah H Collum. 2011. Benefits and drawbacks of electronic health record systems. *Risk management and healthcare policy*, pages 47–55.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI.

NHS. 2023. Purpose of the gp electronic health record. Accessed on March 18, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Raj M Ratwani. 2017. Electronic health records and improved patient care: opportunities for applied psychology. *Current directions in psychological science*, 26(4):359–365.

Marcus Rohrbach, Michael Stark, and Bernt Schiele. 2011. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR 2011*, pages 1641–1648. IEEE.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Yuqi Si, Jingqi Wang, Hua Xu, and Kirk Roberts. 2019. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304.

SNOMED. Snomed international. Accessed on March 27, 2024.

Irena Spasic, Goran Nenadic, and 1 others. 2020. Clinical text data in machine learning: systematic review. *JMIR medical informatics*, 8(3):e17984.

Maryam Tayefi, Phuong Ngo, Taridzo Chomutare, Hercules Dalianis, Elisa Salvi, Andrius Budrionis, and Fred Godtliebsen. 2021. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(6):e1549.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2020a. Neural machine translation with byte-level subwords. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9154–9160.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020b. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

A Appendix

A.1 Examples generated from LLMs

For Experiencer:

- His younger sibling is receiving chemotherapy for colon cancer. They attend oncology visits together; 'colon cancer' - Family
- The physician diagnosed her with Hodgkin Lymphoma during last tuesday's session; 'Hodgkin Lymphoma' - Patient
- The support group aimed at creating awareness among individuals suffering from multiple sclerosis in their community; 'multiple sclerosis' - Other

For Presence:

- At my annual checkup, the GP recommended having a colonoscopy due to family history; 'colonoscopy' - Present

- Patients who have severe kidney damage might require dialysis therapy temporarily or permanently; ‘kidney damage’ - N/A
- Upon reviewing the patient’s file, it appears there have been no diagnoses related to asthma or allergies; ‘asthma’ - Not present

For Temporality:

- Based on current symptoms and test results, the patient will require hip replacement surgery in a couple of months; ‘hip replacement surgery’ - Future
- The patient underwent routine mammography today and has received the imaging results; ‘mammography’ - Recent
- Past X-ray examination indicated signs of osteoporosis, calling for medications and lifestyle changes; ‘osteoporosis’ - Past

A.2 LLM prompts for zero and few shot approaches

A.2.1 Prompt for Mistral 7B instruct model

""" *<s>[INST]You are a text classification bot.*
Your task is to assess intent and categorize the input text into one of the following predefined categories:
2: Experiencer - Patient / default, 1: Experiencer - Family, 0: Not applicable
Explanation of labels: Label 2 (patient / default) is the class where the context strongly indicates that the given medical entity is for the patient. The text will not explicitly contain mention that it is for the patient, you have to infer it. Label 1 (family) is the class where the context clearly indicates that the given medical entity is for the family. Label 0 (not applicable) is when the input data does is not applicable to the category.
You will only respond with the predefined category. Do not provide explanations or notes.
Inquiry: text [/INST] """

A.2.2 Prompt for Llama 3.1 8B instruct model

"""*<|begin_of_text|><|start_head_id|>system <|end_header_id|> You are a text classification bot. Your task is to assess intent and categorize the input text into one of the predefined categories. <|eot_id|> <|start_head_id|> user <|end_header_id|> Classify the input text into one of the following predefined categories:*
2: Experiencer - Patient / default, 1: Experiencer - Family, 0: Not applicable
Explanation of labels: Label 2 (patient / default) is the class where the context strongly indicates that the given medical entity is for the patient. The text will not explicitly contain mention that it is for the patient, you have to infer it. Label 1 (family) is the class where the context clearly indicates that the given medical entity is for the family. Label 0 (not applicable) is when the input data does is not applicable to the category.
You will only respond with the predefined category. Do not provide explanations or notes.
Inquiry: text <|eot_id|> <|start_header_id|> assistant <|end_header_id|> """

A.3 Summary of the modeling approaches employed

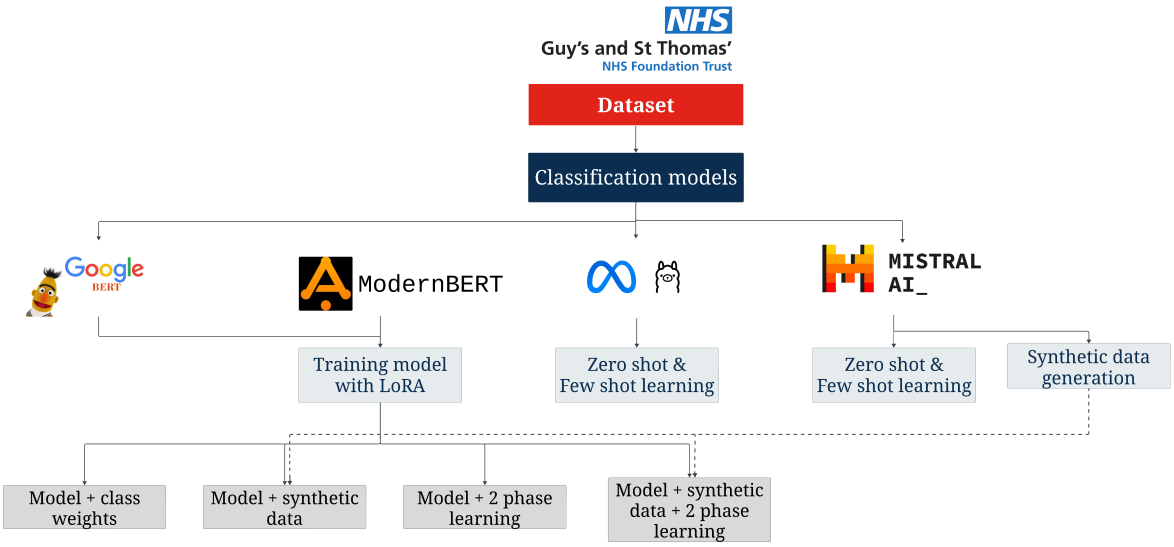

Figure 2: Overview of modelling workflow

A.4 Results from the ablation study across models and tasks

CW - class weights in favour of minority classes; 2PL - 2-phase learning fine-tuning approach + CW; SD - inclusion of synthetically generated data + CW

* indicates the majority class for the task.

Note: The baseline models (models with CW) for Bi-LSTM, BERT and ModernBERT have been fine-tuned on the dataset

Table 3: Model performance for all tasks - ablated

Task	Model	Accuracy	Macro	Recall		
			F1-score	*Not present*	*N/A*	*Present*
Presence	Bi-LSTM (w/ CW)	**0.89**	0.78	0.77	0.72	0.93
	Bi-LSTM (w/ SD)	0.87	0.8	0.79	0.75	0.9
	Bi-LSTM (w/ 2PL)	0.88	0.81	0.76	0.77	0.91
	Bi-LSTM (w/ 2PL + SD)	**0.89**	0.84	0.84	0.79	0.92
	BERT (w/ CW)	0.86	0.82	0.8	0.77	0.91
	BERT (w/ SD)	0.87	0.82	0.8	0.79	0.88
	BERT (w/ 2PL)	0.88	0.85	0.85	0.78	0.91
	BERT (w/ 2PL + SD)	**0.89**	**0.87**	**0.87**	**0.84**	0.9
	ModernBERT (w/ CW)	0.86	0.83	0.83	0.79	0.9
	ModernBERT (w/ 2PL + SD)	**0.89**	0.85	0.86	0.8	0.93
	Llama 3.1 8B (few shot)	0.84	0.45	0.6	0.03	**0.97**
	Mistral 7B (few shot)	0.8	0.38	0.1	0.2	0.95
				Other	***Family***	***Patient***
Experiencer	Bi-LSTM (w/ CW)	0.9	0.77	0.77	0.64	0.92
	Bi-LSTM (w/ SD)	0.91	0.78	0.75	0.68	0.92
	Bi-LSTM (w/ 2PL)	0.92	0.82	0.83	0.7	0.93
	Bi-LSTM (w/ 2PL + SD)	0.92	0.83	0.84	0.73	0.93
	BERT (w/ CW)	0.87	0.84	0.83	0.81	0.9
	BERT (w/ SD)	0.88	0.87	0.84	0.85	0.91
	BERT (w/ 2PL)	0.91	0.87	0.82	0.82	0.94
	BERT (w/ 2PL + SD)	**0.93**	**0.93**	**0.89**	**0.94**	**0.95**
	ModernBERT (w/ CW)	0.9	0.8	0.76	0.78	0.94
	ModernBERT (w/ 2PL + SD)	**0.93**	0.87	0.83	0.84	**0.95**
	Llama 3.1 8B (few shot)	0.69	0.51	0.05	0.9	0.75
	Mistral 7B (few shot)	0.74	0.53	0.17	0.65	0.8
				Past	***Future***	***Recent***
Temporality	Bi-LSTM (w/ CW)	0.87	0.79	0.72	0.78	0.91
	Bi-LSTM (w/ SD)	0.87	0.8	0.75	0.77	0.9
	Bi-LSTM (w/ 2PL)	0.87	0.81	0.74	0.82	0.91
	Bi-LSTM (w/ 2PL + SD)	**0.91**	0.84	0.75	0.84	0.93
	BERT (w/ CW)	0.82	0.8	0.8	0.78	0.83
	BERT (w/ SD)	0.84	0.81	0.79	0.79	0.85
	BERT (w/ 2PL)	0.84	0.84	0.82	0.85	0.85
	BERT (w/ 2PL + SD)	0.87	**0.86**	**0.84**	**0.86**	0.89
	ModernBERT (w/ CW)	0.86	0.8	0.7	0.81	0.91
	ModernBERT (w/ 2PL + SD)	0.92	0.84	0.79	**0.86**	**0.94**
	Llama 3.1 8B (few shot)	0.8	0.43	0.1	0.36	0.9
	Mistral 7B (few shot)	0.77	0.47	0.27	0.55	0.74

A.5 Summary of the existing NLP ecosystem

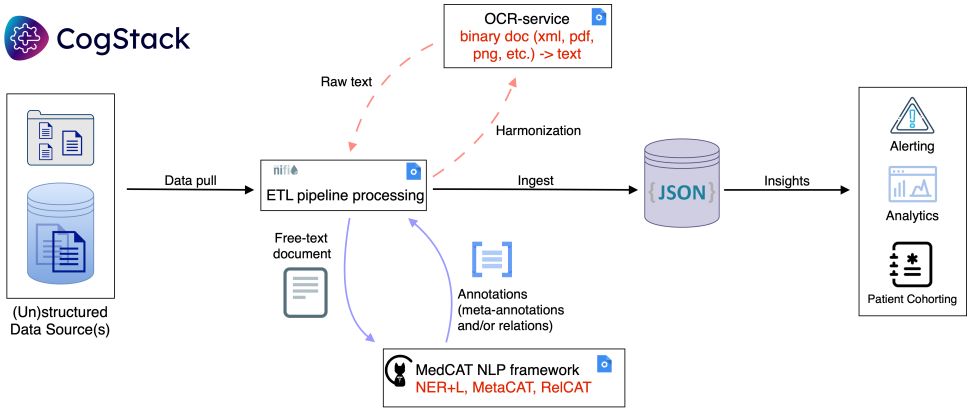

Figure 3: Overview of CogStack ecosystem

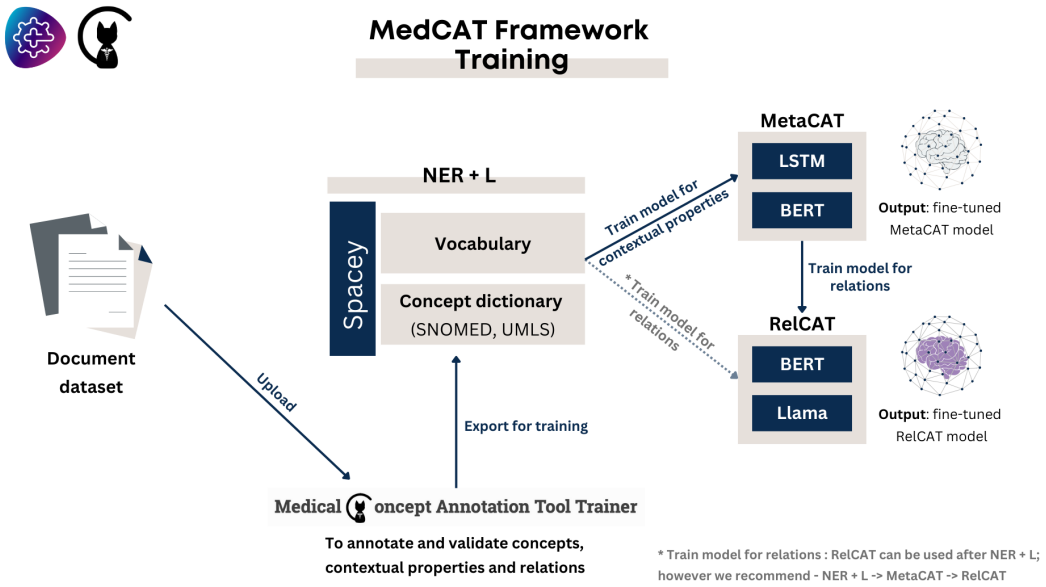

Figure 4: MedCAT framework for training

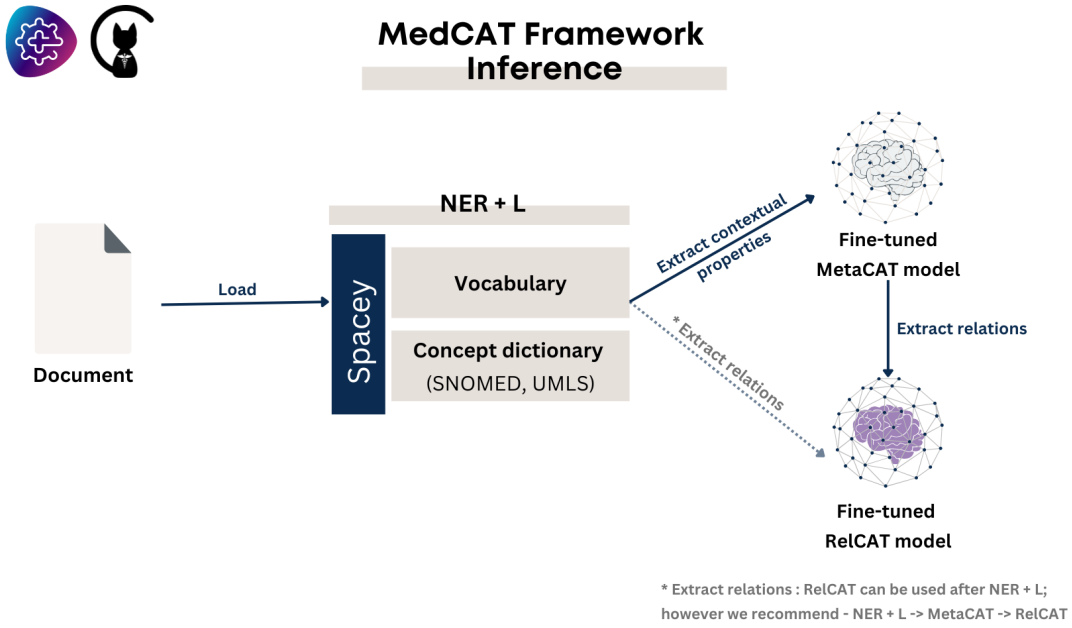

Figure 5: MedCAT framework for inference