A Modern Approach to Adverse Event Detection in Biomedical Corpora

Jonathan Collins, Supervisor: Dr A Altahhan

University of Leeds, School of Computing, ODL MSc in AI, UK.

Abstract. In the pursuit of advancing patient care and medical safety, we present a novel approach to Adverse Event (AE) detection in biomedical corpora, leveraging the latest advancements in Artificial Intelligence. We compare highly optimized classical deep learning techniques using both general and domain-specific BERT-based architectures with the modern approaches of instruction fine-tuning, zero-shot and few-shot prompting. Moreover, we make this comparison across varied biomedical corpora and explore and compare both open-source and closed-source models. We employ Quantized Low Rank Adaptation (QLoRA) for parameter-efficient fine-tuning to address practical considerations such as cost and ethical implications. We conclude by analysing whether these modern techniques do indeed offer improvements that are reproducible in the classical Natural Language Processing problems of Classification and Named Entity Recognition for the task of AE detection. Our results clearly demonstrate that the latest generative biomedical models indeed offer a significant advantage over classical BERT implementations for the detection of adverse events, when they are fine-tuned to specific AE corpora.

Keywords: deep learning, adverse-event detection, parameter-efficient fine-tuning, quantization, low-rank adaption

1 Introduction

The biomedical domain holds a wealth of textual data ranging from clinical notes, research articles and electronic health records to individual patient reports. Much of this data is unstructured or received through a variety of different channels and therefore extracting meaningful information from it is a significant task with implications for patient care, drug development, and the broader field of medical research. Named Entity Recognition (NER) and text classification play critical roles in this context, facilitating the identification of medical terms, drug names, diseases, and various other entities from text, as well as categorizing text segments based upon their content. In the context of this paper, our end goal is to identify Adverse Drug-Events (ADEs, or more simply AEs) within unstructured text, such as below, where the we can clearly see that we have a drug and its adverse effect:

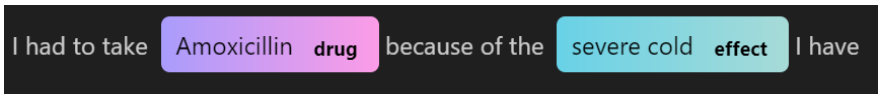

Fig. 1. Example of AE with NER tagging of 'Drug' and 'Effect'

Employing large language models (LLMs) to identify AEs will ensure that important information pertaining to patient safety is not overlooked and is both captured within Pharmacovigilance (PV) databases and subsequently reported to regulators. The identification of AEs using classification and NER, while appearing relatively straightforward, offers challenges owing to the complex and specialized nature of biomedical language, the ambiguity of medical terminology, and the nuanced relationships between biomedical concepts. The influence of social media and medical forums also means the language can be both formal and colloquial, posing additional hurdles to overcome.

Historically, these NLP tasks were performed for AEs using classical machine-learning (ML) models before Transformers were introduced (Vaswani, 2017) was established, significantly outperforming all previous architectures. This led to a plethora of research into transformer-based architectures, including applications

to biomedical corpora like AE data. General-domain pre-trained language models (BERT: (Devlin, 2018), and RoBERTa: (Liu, 2019)) were subsequently outperformed by pre-trained biomedical-specific models like BioBERT (Lee, 2019), SciBERT (Beltagy, 2019), ClinicalBERT (Huang, 2019) and PubMedBER (Gu, 2021), which were considered state-of-the-art (SOTA) for the task AE detection as of the early 2020s, with some incremental improvements of new biomedical-specific BERT models appearing over the past few years.

The advent of Generative Large Language Models (LLMs) has heralded a new era in Natural Language Processing (NLP), demonstrating remarkable capabilities in understanding and generating human-like text: GPT-4 (Open AI, 2023), Llama (Meta AI, 2023) and PaLM (Google, 2022). However it is not well-understood if these advancements translate into more classical NLP tasks such as classification and NER, and if indeed they can offer an improvement on the BERT architectures when applied to the AE detection task. A summary timeline of comparative biomedical-specific (above line) and general-purpose (below line) LLMs can be seen here:

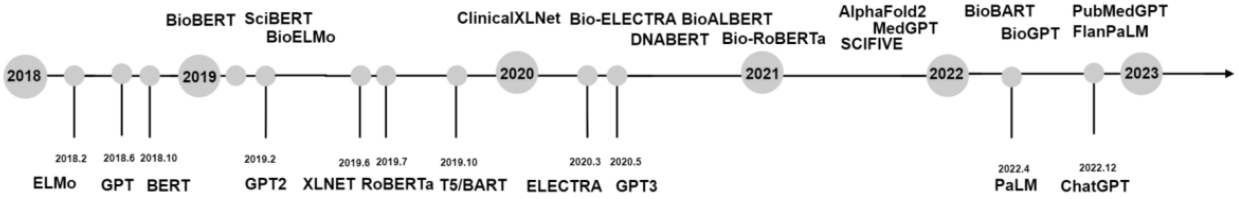

Fig. 2. Timeline of some key BERT and non-BERT models taken from (Wang, 2023)

However we note that even on a daily basis, incrementally better models are appearing. In this paper, we first implement and survey optimized encoder-based BERT models, then move on to specifically compare fine-tuned open-source LLMs and encoder-based Generative pre-trained Transformer (GPT) architectures on the task at hand. The questions remain: Can these more recent and sophisticated LLMs compete in the task of AE detection and bring about a paradigm shift in how biomedical text data is processed and analysed? Or do the older, more established BERT-based models still hold their ground in terms of performance and reliability? This paper provides a comprehensive comparison and rigorous analysis of these different generations of models and further employs the modern parameter-efficient fine-tuning (PEFT) technique QLoRA (Dettmers, 2023) for model quantization to help address practical issues of costs, implementation and speed.

2 Literature Review

The study of AE detection was initially addressed as a classical ML task using Support Vector Machines (Liu, 2013) and Random Forests (Rastegar-Mojarad, 2016) until transformer-based architectures were introduced in 2017. General-domain pre-trained language models BERT and RoBERTa were subsequently employed, however soon outperformed by their biomedical-specific counterparts (e.g. BioBERT, SciBERT, ClinicalBERT and PubMedBERT) which are pre-trained on biomedical-specific data. ScispaCy (Neumann, 2019) is another common out-of-the-box toolkit also employed for NER within pharmaceutical companies for these NER tasks. As of the early 2020s, the use of these domain-specific BERT models and ScispaCy are commonplace, considered state-of-the-art, and are frequently implemented in model pipelines or ensembles to aid with the task of AE detection.

BERT-based models are encoder-based and typically lend themselves to tasks such as classification and NER, especially when fine-tuning is required on a nuanced dataset. A consistent increase in the quantity of biomedical training data, architectures, models and related research publications have pushed the standards yet higher in this domain. With the advent of recent breakthroughs in generative AI the question arises as to whether the encoder-based BERT models still excel in the AE detection task, or whether the likes of GPT-4 or large open-source LLMs have bridged the gap or even overtaken them.

Variations of fine-tuning methods such as PEFT (Parameter-Efficient fine-tuning) have shown to be effective (Fu, 2022), and can be combined with memory-efficiency techniques like Quantized Low-Rank Adaptation (QLoRA). This means billion-parameter models are now accessible to individuals and corporations for domain-specific tuning on modest Graphical processing units (GPUs). The fine-tuning can now be performed quicker, requires less compute power and preserves task performance, therefore it is easier to train and compare more models in a shorter time. However, much recent LLM focus is on Natural language generation (NLG), for example generating dialogue through chatbots like ChatGPT and Bard, rather than more general Natural Language Processing (NLP) tasks such as NER and classification. Therefore we need to understand better whether improvements are reproducible in more classical NLP problems like NER, Classification and Information Extraction (IE).

Existing studies of AE detection in literature often focus on specific subsets of AE data, for example the identification of AEs in tweets (Breden, 2020; Li, 2020; Chen, 2019) or in Clinical notes (Silverman, 2023; Gema, 2023). Tweets and clinical notes are somewhat different due to the use of more colloquial and note-form text, which means that different models, for example those adapted to tweets can perform better than those trained on text from scientific literature publications. Therefore we are concerned more about the problem of identifying AEs in a more general setting with varied drug/event combinations, so we tackle multiple corpora which are implicitly different.

The few literature articles which do exist and address AEs more holistically often lack a clear comparison of modern model performance and fine-tuning approaches. A Biomedical Language Understanding and Reasoning Benchmark (BLURB) was introduced (Gu, 2020) to benchmark the performance of LLMs across a variety of biomedical NLP tasks and can be used as a starting point to identify those which are expected to perform well for a specific task. However the datasets which are used for comparison are often small and nuanced and not domain-specific, therefore a comparison across multiple AE-specific datasets and limited to specific NLP tasks are required for a more decisive comparison.

(Scaboro, 2023) addresses adverse event extraction within two corpora, CADEC and SMM4H. It categorizes the model architectures to three macro-categories: AutoEncoding, AutoRegressive and Text-to-Text. AutoEncoding models are defined as having an architecture composed of a stack of Transformer encoders and contain the BERT models as a subset. AutoRegressive and Text-to-Text models take text input and return text output and also use the decoder transformer architecture. That paper offers an excellent and broad comparison of models from those different architectures, however it does not study models of more than 600 million parameters, whereas we address multi-billion parameter models and a broader range of techniques such as instruction fine-tuning, prompting and quantization.

An excellent survey of the use of LLMs in the Biomedical domain is provided in (Wang, 2023), in which all areas are addressed starting from data collection, through to pre-trained model selection and fine-tuning. It provides clear descriptions, timelines, performance analysis and use-cases for a large variety of model types, including more recent generative GPT-based models. It does not however specifically focus on the task of AE detection, rather it provides a broad survey of use-cases from the biomedical domain, providing a clear landscape which can act as a starting point for our task.

(Gu, 2023) explores the distillation method of using GPT-4 as the teacher and BioGPT as a student model and also looks at one of the corpora we compare, the ADE corpus, and biomedical knowledge extraction. As per that paper, it was noted that GPT models were superior for generative tasks like question answering and summarization, but performed worse with structuring tasks like knowledge extraction and NER. However it is discussed that work is required to explore this area more thoroughly and compare with BERT architectures. It notes that the BERT student model performs better than the generative model when acting as a student, but it concludes that expanding the training corpus and clinical tasks is an area of future exploration. We address this in our study by covering more corpora and looking at both NER and classification, as well as looking across multiple models of different architectures.

The use of parameter-efficient fine-tuning and quantization is not addressed much in biomedical AE literature, mainly due to the fact that biomedical-specific GPT models are very new and older BERT models did not require such methods due to their reduced parameter sizes. (Gema, 2023) uses a LoRA adapter built upon LLAMA and a task-specific adapter which when combined produce SOTA on Clinical NLP tasks, and offer significant computation advantages. The paper noted its limitation about the use of a specific dataset and it is targeted at the clinical domain and not the broader pharacovigilance domain in which adverse events occur.

For instruction fine-tuning, there exist several papers which focus on the use of ChatGPT and OpenAI instruction tuning. Hu (2023) explores clinical NER tasks and the improvement of ChatGPT over previous GPT, however notes that it performs significantly worse than the supervised BioClinicalBERT model when applied in a zero-shot approach. (Kosprdic, 2024) extends this to few-shot and improves upon the zero-shot performance. It notes that is is closer to BERT performance but it still is not on a par. It also notes instability issues. Our paper will investigate whether the newer GPT-4 model provided by Open AI has improved further since this paper, moreover it will address both the NER and more holistically the AE classification task. Additionally, it will extend these studies to compare across multiple corpora and furthermore whether such closed-source LLMs have comparable open-source alternatives, and to what extent they can compete.

To my knowledge, this paper serves as the first study which not only compares multiple AE corpora and provides a comprehensive benchmark of a variety of model performances for Pharmacovigilance but also focuses on modern approaches such as instruction fine-tuning and prompting, incorporating quantization and parameter-efficient techniques which are not well-studied in the domain of AE detection. It also focuses and compares some of the newest biomedically-pretrained generative models which are not well-studied, especially for AE detection. **The aims of this study are threefold**:

- To **analyze** more than one AE corpus obtained across different settings.
- To **compare** AE classification and NER capabilities of:
  - SOTA general and domain-specific BERT-based models.
  - Large open-source LLMs which are both quantized and parameter-efficiently fine-tuned with QLoRA.
  - Large general-purpose Generative LLMs (e.g., GPT-4).
- To assess **performance** metrics, compute resources/time, and practical and ethical considerations.

The study also aims to:

1. **Clarify** whether newer domain-specific BERT models significantly outperform older ones.
2. **Explore** whether recent general-purpose LLMs such as GPT-4 can perform adequately for the AE detection task.
3. **Analyse** the use of fine-tuned open-source LLMs for AE detection, notably whether their performance matches the BERT or GPT-4 models and whether quantized and parameter-efficient implementations are effective and viable in a corporate setting.
4. **Identify** the importance of the source AE corpus and whether this impacts the model choice itself.

We engage with the broader theoretical and practical aspects such as ethical and legal implications of deploying these AI solutions as well as business rationale and implications, understanding that technological advancements must be balanced with ethical considerations, societal impact and corporate viability. This comprehensive survey of literature combined with the results from this paper is not just foundational; it is a strategic exploration aimed at identifying future directions of study, research and the implications of modern AI breakthroughs and techniques for AE detection.

3 Methodology

In this section we provide details of our approach, architectures and data used for AE detection.

3.1 Data

Due to the sensitive nature of AE data, obtaining corpora is itself a challenging task. Three quality AE corpora were identified: **ADE Corpus** (Gurulingappa, 2012), **CADEC** (Karimi, 2015) and **PsyTAR** (Zolnoori, 2019) . These are labelled, annotated and cleaned. Furthermore, they are good representations since the CADEC and PsyTAR corpora are quite different in nature to the ADE Corpus, being sourced from a medical forums, whereas the ADE Corpus is more scientific and formal. Having a significantly different make-up between the data sets (such as some containing slang terms and misspellings) offers the challenges we are looking to address as part of our language modelling task. By applying different techniques and architectures to these opposing datasets, this provides a more holistic view of how both classical and modern

techniques generalise regardless of the corpus itself, and therefore which may be employed more broadly in a corporate business setting.

The ADE Corpus contains both classification and NER labelled AE data sets, PsyTAR is for classification only and CADEC is for NER. We will also address how NER can directly be mapped to an AE detection task in a restricted way. We will address both NER and classification, but ultimately our final goal will be focussed on the classification of AEs in unstructured text and therefore we will pose the overall problem as one of classification, since this is one of the key challenges for PV.

– 1) **ADE corpus** This data set is sourced from 3000 medical case reports. For the biomedical domain, this source is perhaps the most important and relevant since it contains critical information about patients' medical treatments and effects of drugs which can be critical signals for the safety and efficacy of pharmaceutical products and subsequently may be reported to regulatory bodies who maintain oversight of the safety profile of medicines. The ADE corpus is a collection of MEDLINE case reports which are freely available to the public. 5063 drugs and 5776 ADE annotations distributed over 4272 sentences. It consists of 4272 annotated sentences containing 5063 drugs and 5776 ADE annotations. The average sentence length is approximately 20 words. It is focused upon AEs and is partitioned for the two tasks of Classification and NER, with a binary classification output (i.e. '1': has AE, or '0': no AE present), and a separate entity-labelled output (drugs,effects). For the classification task there are 23,516 records with 6821 (29 percent) containing AEs and 16,695 (71 percent) without AEs. For the NER task, there are 6821 records of AEs, for which the drugs and effect locations are indexed within the sentences.

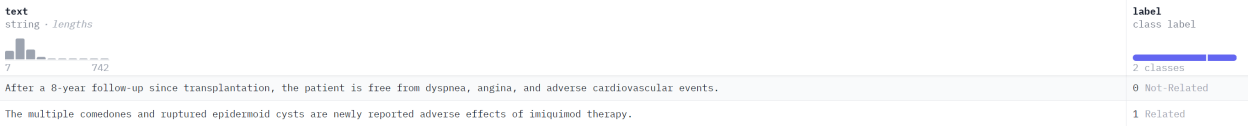

Fig. 3. Positive and negative example from ADE corpus

– 2) **CADEC (CSIRO Adverse Drug Event Corpus)**
CADEC is an annotated corpus of AEs reported by patients through medical forums between 2001 and 2013. It is sourced from 1253 posts from the "askapatient.com" health forum and the grammar is generally colloquial English commonly used in a social media setting. There are a total of 7398 sentences which are annotated with medications, negative drug effects, symptoms, and diseases. The sentences are often long, but can be well-structured and descriptive. There are a total of 1250 forum posts with 1107 posts with at least one AE and 146 with no AEs. Since this is an NER dataset, we are not concerned with this imbalance, only with the present of named entities such as drug and effect. The entities tagged are: ADE, Drug, Disease, Symptom, and Finding. Twelve drugs are included in the reports, of which there are 1800 mentions, and there are a total of 7409 adverse events mentioned. We focus only on the identification of drugs and effects.

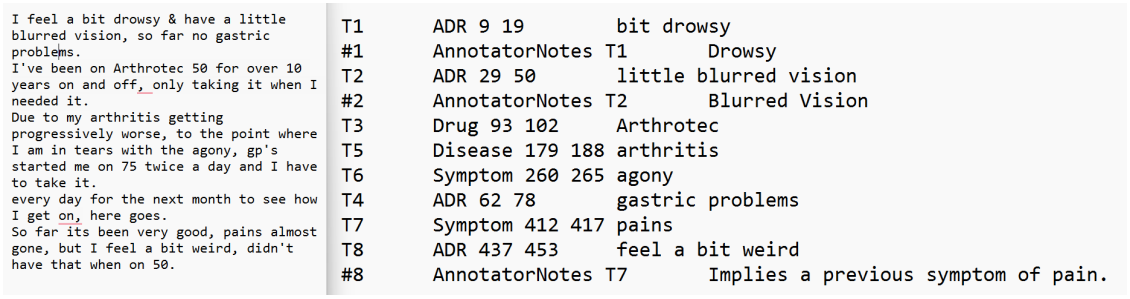

Fig. 4. Example from CADEC corpus: raw text and tagged entities with locations

– 3) **PsyTAR - The Psychiatric Treatment Adverse Reactions** PsyTAR is a corpus sourced from patients' view on the effectiveness and adverse events associated with psychiatric medications. It is sourced from a sample of 891 drug reviews posted on an online health forum by the patients. Four drugs are reviewed (Zoloft, Lexapro, Cymbalta, and Effexor XR). There are 6009 sentences with 4813 adverse events identified. Each sentence is classified as 0 (contains no AE), or 1 (contains AE(s)), making it appropriate for the AE classification task. 2168 of the 6009 sentences contain AEs.

text	label
I slowly cut my dosage over several months and took vitamin supplements to help.	0

Fig. 5. negative example from PsyTAR corpus

It should be noted that the ADE Corpus is sourced from MEDLINE reports and that some MEDLINE data is included in the PUBMED pre-training of some BERT and generative models for which we are studying, however we will consider this when comparing the metrics and it should be noted that our fine-tuned models based upon our datasets for the specific NER and Classification tasks could regardless be highly powerful when predicting on unseen data.

Data Preprocessing: Biomedical NER (Bio-NER) is a critical task in performing information detection and extraction for clinical decision-making tasks. Labelled data in the biomedical domain can be challenging to collect and the entities can be complex and different to other fields of research. For the NER models we prepare the data using the standard CoNLL format and use BIO (Beginning, Inside, Outside) tagging method which is the most commonly used in the biomedical domain. The final prepared data has a labelled list pair: text: ['stomach', 'pain', '.'], labels: ['B-ADR', 'I-ADR', 'O']

Each sentence is taken from the corpus and tokenized, then labelled with BIO-tagging based upon the type and position of the tokens inside entities. Special CLS, SEP and PADDING tokens are used to prepare the text. These tokenized inputs are passed to the model, with a linear layer applied to predict the token tags.

We limit the high-level NER entity categories to 'drug' and 'effect' (marked as 'ADR' in the below diagram) in both the NER datasets (CADEC and ADE Corpus) to have a comparable task.

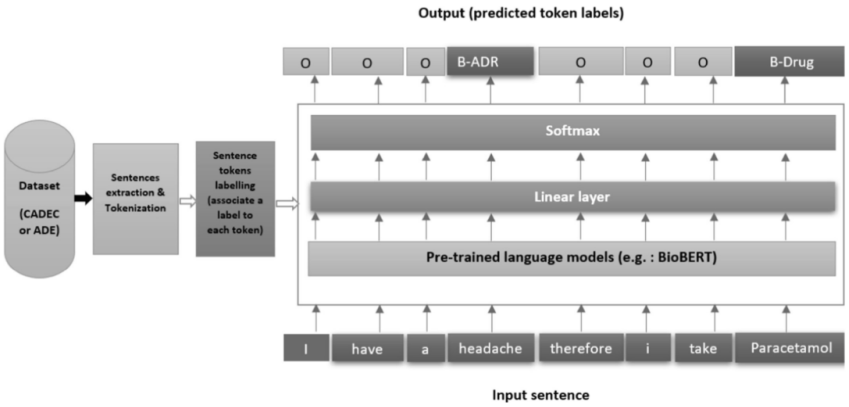

Fig. 6. NER token classification model end-to-end high-level, diagram from (Hiba, 2023)

For the classification models the datasets are already clean. They are all labelled and ready for the supervised classification AE detection task. There is a distinction between the target class of the two classi-

fication datasets: The ADE corpus looks for a 'related' adverse drug-event, i.e. one that could be specifically attributed to a drug ('1' for those where a possible drug side effect is present in the text, 0 otherwise); The PsyTAR corpus is taken from a drug discussion setting, therefore assumes the presence of a drug and looks simply for the adverse drug effect in the sentence ('1' if present, 0 otherwise). The slight difference in targets will play an important role in telling us which model architecture generalises better to the AE classification task.

In all instances, the data is split into a training and test set, with the test set put aside for the metrics comparison of the models on unseen data.

3.2 Model Architectures

Here we describe the architecture and models used in our research.

3.2.1: BERT-based models

All BERT-based models are pre-trained and sourced from the HuggingFace library (HuggingFace, at 04Mar024). They include one two non-domain-specific models, BERT-large and SpanBERT, and six domain-specific models. BioLinkBERT, BioBERT v1.1 and v1.2, EnDR-BERT, BioClinicalBERT and SciBERT

(a) Non domain-specific BERT models

-**BERT**: (Bidirectional Encoder Representations from Transformers) is trained with bi-directional transformers from large amounts of unlabelled text and uses two training objectives: masked language modelling and next sentence prediction. Due to complicated interactions in biomedical terminology, BERT's bidirectional capabilities excel. The model size is 109 million parameters.

-**SpanBERT** Is pre-trained for improving the representations and prediction of text spans. The masking mechanism differs from BERT in that it masks spans of consecutive tokens instead of random ones. It employs a span-boundary objective to help predict the entire masked span based on boundary tokens. The model size is 108 million parameters.

(b) Domain-specific BERT models:

-**BioBERT**: One of the first Biomedically pre-trained BERT models. v1.2 is the updated version from v1.1 with a slightly improved performance. The biomedical corpora in training included texts from PubMed abstracts and PMC full-text articles. The model size is 109 million parameters. The model size is 109 million parameters.

-**SciBERT**: Was trained from a sample of over 1 million documents from Semantic Scholar and contains a mix of biomedical and computer science publications. Entire texts were used, and not just. It was trained on over 3 billion tokens, similar to the initial BERT model. The model size is 109 million parameters. The model size is 109 million parameters.

-**BioLinkBERT** uses information from biomedical knowledge graphs, which are structured representations of knowledge found in the biomedical domain. This allows the grasp of complex relationships between biomedical entities such as genes, diseases, and drugs more effectively than text-only trained models. The model size is 340 million parameters.

-**EnDR-BERT** is a multilingual and trained consumer comments on drug administration. It is similar to Multi-BERT and another variant upon the initial BERT model. The data contains user-generated texts from multiple internet sources which include many social media posts. The model size is 177 million parameters.

-**BioClinicalBERT** is specifically trained on clinical text and therefore may perform well when applied to data from clinical trials or in a clinical setting. It was trained on a large corpus including electronic health records, clinical trial reports, and other medical documents. The model size is 108 million parameters.

Each BERT-specific model has a slight nuance, but the general architecture used for pre-training and fine-tuning is as shown in figure 7. Fine-tuning is quite simple and uses the self-attention transformer mechanism to allow the BERT to be applied to downstream tasks, by adapting the inputs and/or outputs. For each task, we simply plug in the task-specific inputs and outputs into BERT and fine-tune all the parameters end-to-end. At the output, we add a classification head to provide the appropriate binary output.

Once the data is prepared, tokenized, for the classification task, a classification head is added with a dense layer for binary classification. For NER, a token classification head with an output size equal to the number of unique entity labels is added. Data is prepared with dataloaders and appropriate loss function,

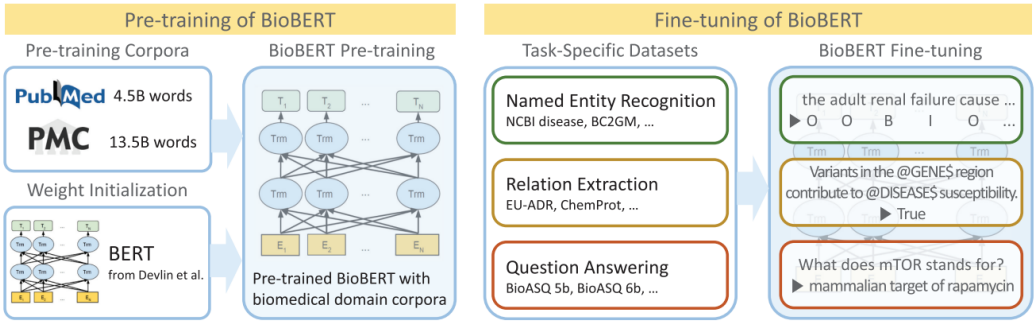

Fig. 7. Pre-training and fine-tuning of BERT-based models, taken from (Lee, 2020)

optimizer, learning rate scheduler and training loop employed. Hyperparameter sweeps were employed to find the best parameters during training, looking at both how the training and validation sets performed. A custom pytorch implementation was coded and evaluation was performed from the loss function results and using precision, recall and f1 score on the validation set. Each model was fine-tuned on each of the relevant datasets to provide optimal performance (ADE corpus and PsyTAR for classification; CADEC and ADE corpus for NER).

3.2.2: Fine-tuned Open-Source models

The performance of open-source models is drawing attention as they compete and even surpass some closed-source models. Open source models offer significant benefits especially in the healthcare sector where privacy can be an issue and data cannot be shared. Therefore such models can offer significant advantage, if they can perform well, they can be housed in-company. A number of biomedical-specific generative models are now available which are pre-trained on biomedical text and can be fine-tuned on task-specific data. We compare the performance of general domain models OPT-2.7 (Zhang, 2022) and Mistral-7B (Jiang, 2023) with domain-specific pre-trained models BioMedLM (Bolton, 2022), BioGPT (Luo, 2022), BioMedGPT (Zhang, 2023), BioMistral (Labrak, 2024).

(a) Non-Domain specific generative models:

-**OPT2.7** Based upon Open Pre-trained Transformers (OPT), this model is decoder-only and similar to GPT architectures, with a performance similar to GPT-3 models. It is mainly pretrained with English text, using a causal language modelling objective.

-**Mistral-7B** Mistral's 7–billion-parameter model outperformed larger models such as Llama 2 across all evaluated benchmarks, and even large 30GB+ models like Llama 1 in several tasks. It uses grouped-query attention for faster inference and a sliding window attention to deal with abritrarily long sequences without significant inference impact. We use a Mistral-7B-Instruct implementation

(b) Biomedical-specific generative models

-**BioMedLM** is a biomedical-specific generative decoder-only transformer model with 2.7 billion parameters and has a vocabulary size of 28896. It is trained on PubMed abstracts and PubMed Central portions of the Pile dataset, which itself has 50 billion tokens, 16 million abstracts and 5 million full-text articles from the biomedical domain.

-**BioGPT** The BioGPT model is a biomedical-specific GPT model for biomedical text generation and mining. It is pre-trained on 15 million PubMed abstracts with a causal language modelling objective, therefore should be well-suited for our proposed tasks. It is a recent model which has achieved STA results on four benchmarks: BC5CDR, KD-DTI and DDI end-to-end relation extraction task, and the PubMedQA question answering task.

-**BioMedGPT** This model is based on Llama2, and the first specifically of its type for the biomedical domain. It was fine-tuned from the Llama2-7B-Chat on millions of biomedical papers which were collected from the S2ORC corpus. It is an open-source multimodal GPT for biomedicine is yet not well-studied for AE detection.

-**BioMistral** This is an open-source LLM for the medical domain, built upon Mistral models and based upon PubMed Central corpora. Since the Mistral models are some of the top performing open-sourced models, we are eager to test their biomedical-specific implementation.

For the open-source models we provide two different implementations for each model and corpus combination, an instruction fine-tuned causal language model and a fine-tuned sequence classification model. Both implementations use the huggingface transformers library and huggingface trainer. Both are implemented with QLoRA to provide quantization and parameter-efficient fine-tuning. LoRA injects a small number of additional trainable parameters while freezing the other model weights and offers significant benefits for training and inference time. The quantization of QLoRA means that we have a 4-bit implementation using the nf4 quantization, implemented with the bitsandbytes library.

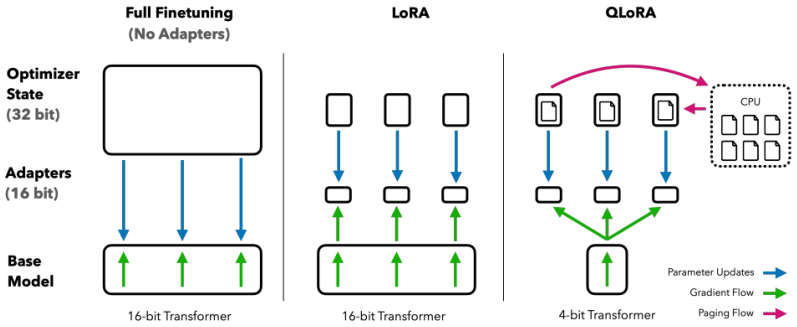

Fig. 8. QLoRA architecture, taken from (Dettmers, 2023)

QLoRA offers an improvement upon LoRA by not only injecting additional weights for the fine-tuning and freezing the model weights, therefore only requiring a small update in the number of model parameters (typically less than 1 percent), but also it adds a quantization improvement which means the GPU requirements are lower. The QLoRA paper shows that the effect of quantization and parameter-efficient fine-tuning does not have a significant effect when compared with full fine-tuning, so this is both an efficient and effective technique.

For The instruction fine-tuning model is given the instruction to return a binary value of 0 or 1, based upon whether the given text contains an AE. Additional instructions are given to ensure no additional output is provided.

3.2.3: Closed-Source models

Open AI's API is used to compare the performance of GPT-3.5 and GPT-4 on the task of AE detection, via binary classification. It is specifically of interest to compare this with the closed-source generative models, especially those pre-trained on biomedical corpora, but also on the other architectures such as the biomedical BERT implementations.

The method used will both zero-shot - purely providing a task for the Open AI's GPT models to classify the text as containing an AE or not, and also few-shot, where we provide some examples and see if the performance improves. This will follow a similar instruction fine-tuning method, but require custom code to communicate with the Open AI API. We will also note the performance across the datasets and comparison with other models in this respect, and how the number of samples in the few-shot instance has an impact. Example prompts can be seen in Figure 9.

3.3 General Approach

Performance Metrics: We use a similar method to other biomedical literature and present the **precision** and **recall** score metrics for the classification task. For the NER task, we provide the micro-averaged class scores. We will display the cross-architecture comparisons to see which type of models excel. We will

```
### Instruction: The input is a list of words and your task is to return a list where where each word has a corresonding named entity.
    These are all the names for each entity class:
    O = None
    B-drug = The word is the drug itself, or first word in part of a drug name
    I-drug = Part of a drug name, but not the first part, i.e. immediatedly following another word which is also part of the same drug
    B-effect = The word is the side effect itself, or first word in part of an effect name
    I-effect = Part of an effect name, but not the first part, i.e. immediatedly following another word which is also part of the same effect

    Guidelines:
    Only output the list of entities and nothing else
    Do not output anything other than a list of the entity names
    For each input list make sure to output a list of the same length
    The output list must have the entity class of each word in the input list only
    You must output a list of length {len(item)}. Each item of the list should correspond to the entity class of each element in the input list.
examples\n

Input : ['Intravenous', 'azithromycin', '-', 'induced', 'ototoxicity', '.']\n
Response : ['O', 'B-drug', 'O', 'O', 'B-effect', 'O']\n

Input : ['Unaccountable', 'severe', 'hypercalcemia', 'in', 'a', 'patient', 'treated', 'for', 'hypoparathyroidism', 'with', 'dihydrotachysterol', '.']\n
Response : ['O', 'O', 'B-effect', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-drug', 'O']\n

Input : ['METHODS', ':', 'We', 'report', 'two', 'cases', 'of', 'pseudoporphyria', 'caused', 'by', 'naproxen', 'and', 'oxaprozin', '.']\n
Response : ['O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-effect', 'O', 'O', 'B-drug', 'O', 'B-drug', 'O']\n

Input : ['Marked', 'elevation', 'of', 'serum', 'creatine', 'kinase', 'associated', 'with', 'olanzapine', 'therapy', '.']
Response : ['O', 'B-effect', 'I-effect', 'I-effect', 'I-effect', 'I-effect', 'O', 'O', 'B-drug', 'O', 'O']
```

Fig. 9. OpenAI GPT-4 Instruction for NER

compare discriminative BERT-based models with their generative adversaries like GPT-based models. Ultimately, for the task of AE detection, we are looking at a good balance of precision and recall, however we are especially concerned with a high recall, as we want to detect AEs, but also minimizing false negatives (missing an AE completely). For this reason we will look for those models with a strong balance across all metrics but mainly one which achieves both a high precision and recall. The performance metrics presented will be based upon the optimized fine-tuned models, either those with the optimal sweep parameters, or those instruction or otherwise fine-tuned for the specific model, task and dataset combination.

Model training: For the BERT-based models, weights and biases *(https://wandb.ai/)* was incorporated into the custom pytorch training code and a combination of bayesian, random and grid-search sweeps were used to find optimized parameter ranges for each model and dataset combination. The input sequence length for the models was fixed to 512 in all instances. For the generative models, some exploration was performed with learning rates, epochs, temperature and other hyperparameters (see later). Also, similar tests were performed with QLoRA to find optimal values for the test and to ensure the number of trainable parameters injected through the LoRA process was low.

For the final evaluation of all models, they were tested on the unseen test set, having chosen the best hyperparameters found during the training process, after tuning on the validation portion. For each task a consistent GPU was utilized. For the BERT models, a GeForce RTX4090 (laptop version) with 16GB or VRAM was used. For the generative models (except Open AI), an **A100** with 80GB VRAM was utilized. The training time varied significantly based upon the model architecture, task and dataset. The initial BERT sweeps took up to 10 hours each, but the decisive fine-tune based on optimal parameters took anything from 2 minutes upwards, with an average of 5 minutes for an NER fine-tune and 27 minutes for the classification task. Due to the size of the ADE corpus, it took an average of three times longer for the fine-tune than the PsyTAR corpus.

For the generative sequence classification models, the A100 implementation was used, and the average training time was 47 minutes, with the PsyTAR again being about three times faster than the ADE corpus. For the causal language models, the average training time was significantly longer at around 3.5 hours. When communicating with OpenAI's API, the response time for the entire classification test predictions was around 30 minutes on average, and around three times quicker on the PsyTAR dataset.

4 Experiment Results

4.1 Named Entity Recognition for Adverse Events

As we previously noted, NER can often be the starting point for classifying medical terminology. It can also be useful when there are misspellings or ambiguous terms which might not be recognised either by humans

or by other automated systems like pattern-matching algorithms. In this paper, we have two NER datasets for which we present results: CADEC and ADE Corpus. Indeed, some institutions may decide not to use a full AE classifier and simply take an NER-tagger to identify if an unstructured text contains either only adverse effects, only drugs or a combination of both, and then provide these to human or other pipelined systems to further analyse this data. For that reason we present NER as an initial step for ADE classification and provide the micro-average metrics across drug and effect tagging in our CADEC and ADE corpora.

Our main goal was to compare the performance of GPT-4 with domain-adapted BERT models to see if they can compete with the industry standard. We optimized both methods to give a clear comparison.

Hyperparameter optimization: The search strategy employed for the BERT models was focused on five hyperparameters: **learning rate, batch size, dropout, weight decay, and epochs**. Some trial and error was involved at first and sweeps were performed with random searches to find useful ranges, in addition to consulting other literature articles for guidance. Once reasonable ranges were established, a bayesian and/or grid search was employed for each combination of model and dataset applicable to the task (for each BERT model combined with either ADE Corpus or CADEC, the best hyperparameters were identified). This search was performed by integrating the custom Pytorch implementation with weights and biases.

The **learning rate** search range was limited to the set [**1e-5,5e-5,1e-6,5e-5**] and a learning rate scheduler employed. Generally a lower starting learning rate performed better. The **batch size** search range was limited to the set [**4,8,12,16**], with a few of the largest models such as 'bert-large' limited to a batch size of 8 as the 3090 GPU ran out of space. Higher batch sizes generally yielded better performing models. The **dropout** search range was limited to the set [**0.1,0.15,0.2,0.25,0.3**]. Generally a dropout of around 0.15 or 0.2 helped to avoid overfitting, especially as the number of epochs increased. The number of **epochs** was kept fairly flexible but most models performed best **between 5 and 10** epochs, mainly on the lower side. If models kept performing well on the validation set, then a longer number of epochs was preferred. The **weight decay** was limited to the set [**0.001,0.01,0.1**]. It generally did not have too much effect, most likely due to the dropout already acting as a regularisation technique to avoid overfitting, but it did seem to bring some small improvements.

For the GPT-4 NER searches, the instruction as per earlier diagram was also changed but it did not make a significant difference. For NER, **zero-shot did not perform well at all**, so few-shot was tried with **varying number of examples provided (maximum 10).** The **temperature** was varied from 0 to 0.5, with **0.1 giving the best results**. GPT-3.5 was also tried but performed very poorly. Overall, the results for GPT-4 were **a little inconsistent,** with the same requests sometime producing varying results by more than a few percent, therefore it was somewhat unstable. Since it is preferred for generative tasks, this is not a complete surprise.

We present the NER results in Figure 10. We compare two architectures, namely fine-tuned BERT implementations (domain and non-domain-specific) and Open AI's GPT-based models. Based upon initial testing, NER performed very badly on a zero-shot approach, with almost no test examples correct, so for the purpose of NER we only consider the few-shot approach, since this is a more complex task where some examples are useful, in contrast to the AE classification task where we also try the zero-shot approach.

For the **few shot** approach, **3** samples were tried at first, and then **10** in total, with slight increases in performance until that point, then similar results. The cost of training with GPT-4 is high, so trials were limited. GPT 3.5 was also tried but the performance was also very bad, significantly worse than GPT-4.

Comparing the BERT models directly, we see that SciBert, BioBERT and BioLinkBert performed best on the ADE Corpus. BioLinkBERT as noted earlier is trained on a much higher number of tokens (300 million+) compared to most other BERT models which have nearer 100 million, so it is of no surprise the performance is strong. Also it is the newest biomedical BERT model tested here. We again note as mentioned earlier that is possible there is bias as some of the ADE Corpus may be included in the pre-training data, but we noted such a fine-tuned model could itself still be very powerful for future prediction.

For the CADEC corpus, we see a very interesting result that the **ENDr-BERT was the best performing**, and most notably that is pre-trained on a significant amount of internet and social media posts. Since the CADEC corpus is of a colloquial nature and from a forum setting, this shows how important the pre-training domain is and how a fine-tuned model can be very effective on the correct data.

It is clear that the BERT-based models significantly outperformed GPT models on the NER task based upon the graphs. Indeed this was expected as the bi-directional encoders used by BERT are

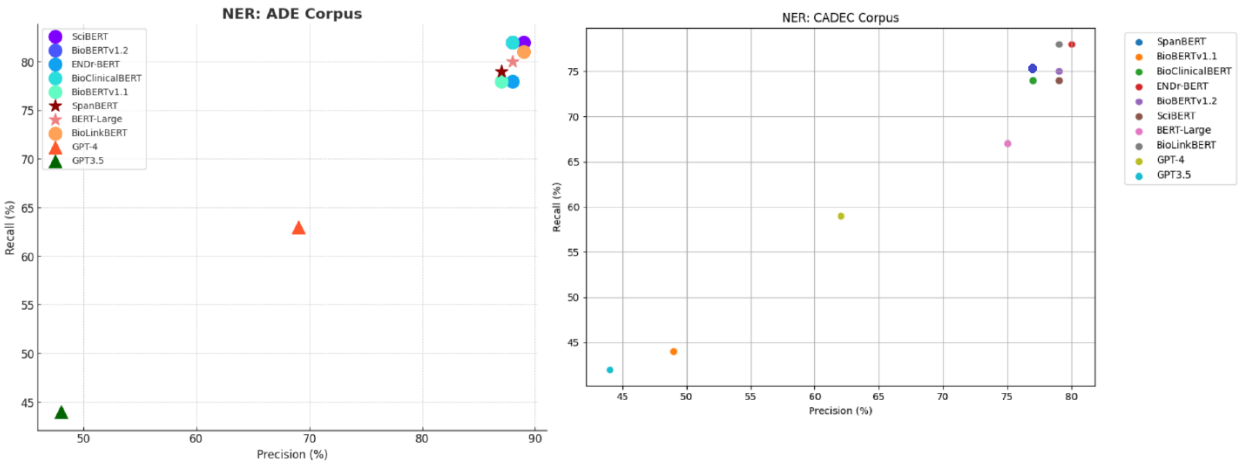

Fig. 10. NER results for ADE and CADEC corpus for BERT and GPT-models

perfect for the NER task. Also, GPT models excel at generative tasks, and the NER task does not align with this. **We can conclude that BERT models are the preferred architecture for the NER task at the moment, in comparison to Open AI's GPT models**. Also, it should be noticed that GPT-4 produced inconsistent results, for example when the same example was given multiple times, it could produce different results. A temperature value of 0.1 was identified as a good value (higher temperature means more creative behaviour), and a value of 0 didn't produce any improvement, whereas higher values such as 0.2 were slightly worse.

4.2 Adverse Event Classification

Our end goal is to classify AEs, in a specific way based upon our corpus. For the AE classification task, we considered a very broad range of models, from the well-tested BERT (general and biomedical-specific), through to instruction fine-tuned causal language models, generative sequence classification models, and Open AI's GPT-4 model. We considered open- and closed-source models, very recent biomedical-generative models and we also quantized and used low rank adaptation for the fine-tuning on the open-source generative models to allow for cost and time-savings.

The hyperparameter search strategy for the BERT models was the same as for NER, with the search strategy focused on five hyperparameters: **learning rate, batch size, dropout, weight decay, and epochs**. The **same** ranges were employed with the same rationale. For the generative models, during the training, the validation set was used to check when overfitting occurred. The number of **epochs** was generally between 2 and 5. The other hyperparameters are addressed in the following section about QLoRA.

For the comparison and graphing of results, we make the distinction BioMistral (S) and BioMistral (C) for each generative model to distinguish between the finetuned Sequence Classification instance and the finetuned Causal Language model instance, on the graph.

Our first observation is that **we can clearly see that the BERT-based models were outperformed by their more modern adversaries in the generative models**. Looking first at the PsyTAR dataset in **Figure 11**, which is from a medical forum and colloquial setting, of the 23 plots on the chart, we see 5 of the lowest 7 recalls were BERT models. However, we again see a strong result for ENDr-BERT, which is the best performing BERT model, having a strong precision at 84 percent and nearly 92 percent recall. Again this echoes how the **pretraining on social media has a huge effect, especially when further fine-tuned as we have for this corpus**.

We see very high recalls for GPT-4, and also OPT-2.7 (both 92 percent and over, by far the best). We note that OPT (S) is similar in architecture to GPT. However, their precision of those two falls significantly lower than the biomedical-specific generative models of BioMedGPT and BioMistral. For the detection of AEs, a high recall is indeed very important, but also a balance of precision is required and we need to reduce the number of false negatives, i.e. we need to avoid missing many AEs. More broadly, the graph shows

that the causal LMs slightly outperform the sequence classification generative models, however it is model dependent.

Looking at the ADE Corpus in **Figure 12**, we see the BERT models performed much better than they did for the PsyTAR corpus. As we noted earlier, **there may be bias** in that this corpus may be part of the pretraining data, so we are cautious in that they may have overestimated performance-wise, but we also note that the bio-generative models also were similarly pretrained and may have similar bias. **Nevertheless, looking at both the precision and recall we see the Biomedical-specific generative models perform best**, with BioMedLM, BioMedGPT and Biomistral towards the top right quadrant. We should also note they are the **generative sequence classification models and we can clearly see those outperform the causal implementations**. With GPT-4, we see a significant difference, in that they don't compete well on this corpus. This is likely to be with the fact that it includes many highly scientific terms which may be found only or more commonly in a biomedical-specific corpus, and therefore may not be represented well in the GPT-4's pre-training data.

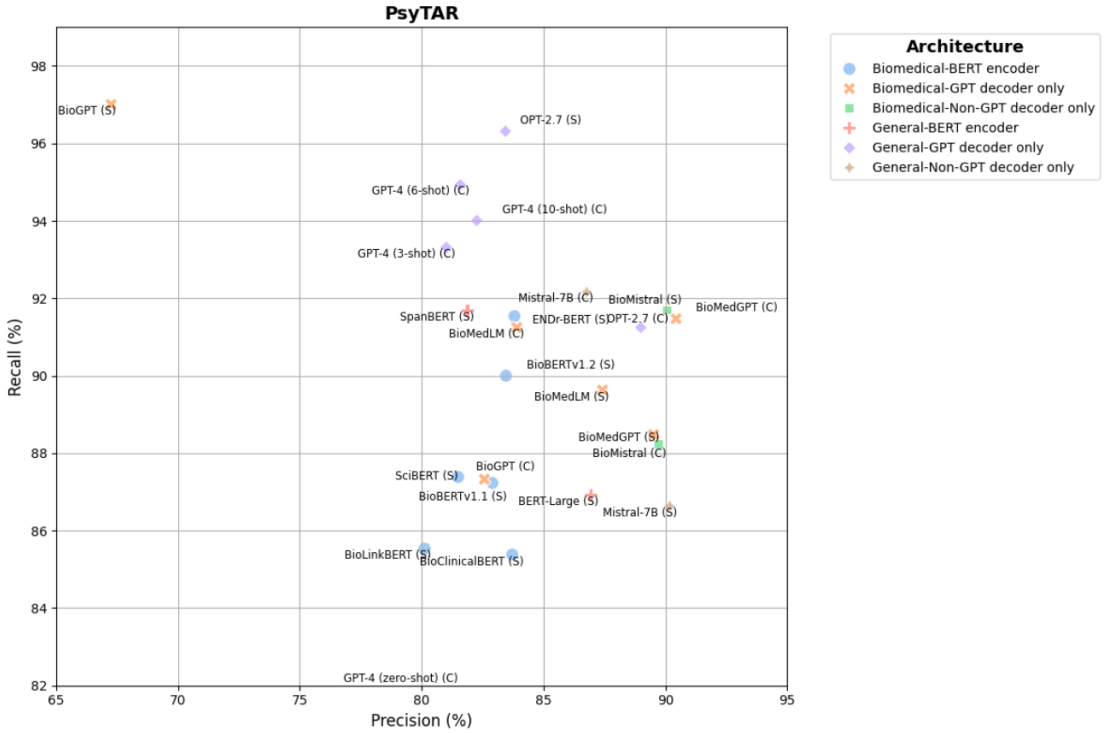

Fig. 11. AE Classification on PsyTAR corpus

4.3 Quantization and Low Rank Adaptation

QLoRA was employed where a **4-bit nf4 quantization** was employed. The number of trainable parameters injected by LoRA was between 0.09 percent and 0.29 percent, with an average of 0.19 percent. For example for the OPT-2.7b model, we had: Trainable params: 2621440 —— All params: 1395927040 —— **Trainable percentage: 0.19 percent.**

For the QLoRA target modules we experimented and generally found that applying to all linear layers including fully connected, performed worse than just the 'proj' modules. A common subset was [**"q-proj","v-proj","k-proj","o-proj"**] or the equivalent modules, which seemed to perform best based upon some trial and error. We explored the **'r'** and **'alpha'** values as hyperparameters. The value of 'r' corresponds to the number of parameters in the adaptation layers, for more complex tasks a higher values is better but it can

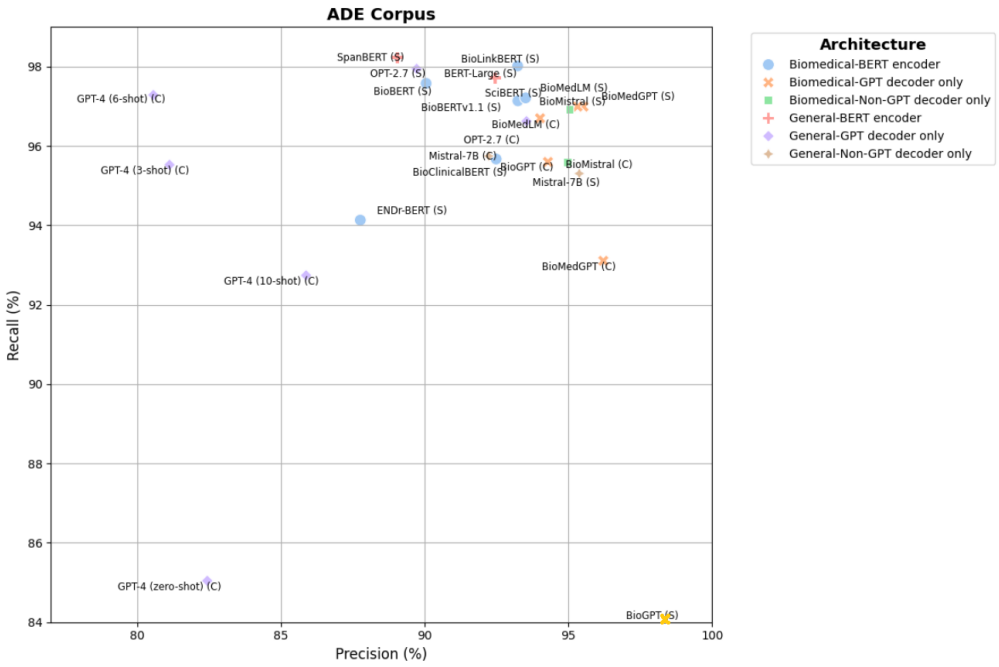

Fig. 12. AE Classification on ADE corpus

also overfit, and here the increase did not improve performance further. Alpha is the scaling factor which how the adaptation layer's weights impact those of the base model, with higher alpha meaning the LoRA layers have more impact on the base model. We identified that **r=4 and alpha = 8** performed generally best across all models, so kept this fixed. Based upon empirical evidence (Dettmers, 2023 and others), an alpha value of twice 'r' works well in practice and we confirmed this with our choice.

Based upon some tests, the training and inference time appeared reduced to full fine-tuning and the GPU usage was significantly reduced. Although tests were performed with an A100, a few trials were performed on the 3090 to check the 7 billion models would fit within the 16GB GPU VRAM, which was concerned. Further testing needs to be performed to effectively measure and assess the extent of the time and cost-saving for the QLoRA technique, however it appears to particularly benefit on the computation requirements of the GPU when moved to consumer hardware, as evidence by fitting to the consumer GPU. The LoRA dropout parameter was also tested as a hyperparameter, with various tests from 0 up to 0.3. The **dropout value of 0.2** was consistently found to perform best to avoid overfitting, hence was applied to all generative models using QLoRA.

5 Conclusion and Future Work

We have conducted a study across multiple datasets for both tasks of NER and ultimately AE detection and classification. We have studied a variety of different models, architectures and performed extensive optimization techniques to obtain the most effective and efficient fine-tuned models available for the biomedical domain. We have employed parameter-efficient and quantized low-rank adaptation with QLoRA, and we have analysed both open- and closed-source solutions, with an aim to having more flexible and customizable solutions.

In summary, we conclude that **BERT-based architectures still excel at the task of named entity recognition**, and GPT-4 cannot compare on this task. Due to the bi-directional encoders employed, this task lends itself specifically to that type of model, whereas GPT-4 is focused more on generative and creative outputs. We find that some of the newer BERT models like BioLinkBERT which have more parameters and a larger biomedical pretraining corpus, offer advantages over many of the older BERT models. We furthermore note that the ENDr-BERT which was trained on a significant portion of social media data, performed very

well (and significantly better than most other BERT modelds), when fine-tuned on the CADEC corpus which comes from a forum setting with more colloquial language. This is quite logical, but we can clearly see that the choice of corpus for fine-tuning and the pre-trained model which is adapted is of utmost importance.

For AE classification, however, we can make much bolder claims that the very **new biomedical domain-specific generative models like BioMistral and BioMedGPT significantly outperform both the BERT models and the closed source models of GPT-3.5 and GPT-4**. Furthermore, those are open-sourced models and could be employed in-house, without private company data needing to leave an organization, whereas for example with GPT-4, you need to send data through the Open AI API, which is not possible for pharmaceutical and other medical companies due to privacy laws and patient confidentiality, especially due to the sensitive nature of adverse drug effects that patients may encounter.

The implementations using QLoRA can further allow cost reduction when training and performing inference on biomedical data such as AE classification. Our fine-tuned models and code are available (Collins, 2024) including full details of optimal hyperparameters for further investigation and inference. It should be noted that Adverse event data is very sensitive and this is why not many public corpora are available, and why testing with closed-source models which require private data to be sent outside of companies is not possible. However, if the closed-source fine-tuned implementation or code from this paper are used, they could further show promising results when applied to larger corpora, in which performance could be even better. The combination of multiple corpora from different settings, or larger private corpora is an area where we could see those significant improvements. In terms of the parameter-efficient fine-tuning and quantization, we have successfully implemented and deployed QLoRA with excellent results, however a more thorough comparison would be required to quantify better the benefits of computation, time and eventually cost which these methods can bring.

We can conclude with a very bright outlook upon these new biomedically-pretrained generative models and expect these to be employed more widely in the pharmaceutical industry, especially for AE detection. Due to regulatory requirements and the impact of missing AEs, it is likely they will not be applied 'out-of-the-box', however with further studies, it is possible we could see them employed in company pipelines to help detect AEs. It is not inconceivable that with more fine-tuning attempts that they could even perform on a par or even surpass human levels of detection, not only offering huge cost savings by allowing these tasks to be automated and performed algorithmically, but ultimately helping to address a key area of patient safety.

6 References

- Ashish Vaswani (2017) "Attention Is All You Need" https://arxiv.org/pdf/1706.03762.pdf
- Devlin (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding arXiv:1810.04805
- Liu (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach arXiv:1907.11692
- Lee (2019) BioBERT: a pre-trained biomedical language representation model for biomedical text mining arXiv:1901.08746
- Beltagy (2019) SciBERT: A Pretrained Language Model for Scientific Text arXiv:1903.10676
- Huang (2019) ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission arXiv:1904.05342
- Gu (2021) Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing arXiv:2007.15779v6
- Open AI (2023) "GPT-4 Technical Report" https://arxiv.org/pdf/2303.08774.pdf
- Meta AI (2023) "LLaMA: Open and Efficient Foundation Language Models" https://arxiv.org/pdf/2302.13971.pdf
- Google (2023) "PaLM 2 Technical Report" https://arxiv.org/pdf/2305.10403.pdf
- Wang (2023) Pre-trained Language Models in Biomedical Domain: A Systematic Survey
- Dettmers (2023) "QLoRA: Efficient Finetuning of Quantized LLMs" https://arxiv.org/pdf/2305.14314.pdf
- X. Liu and H. Chen (2013), "Azdrugminer: an information extraction system for mining patient-reported adverse drug events in online patient forums," in International conference on smart health, pp. 134–150, Beijing, China
- Rastegar-Mojarad (2016) Detecting signals in noisy data-can ensemble classifiers help identify adverse drug reaction in tweets," Proceedings of the Social Media Mining Shared Task Workshop at the Pacific Symposium on Biocomputing, Kohala Coast, United States
- Neumann (2019) ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing arXiv:1902.07669
- Open AI (2023) "GPT-4 Technical Report" https://arxiv.org/pdf/2303.08774.pdf
- Fu (2022) "On the Effectiveness of Parameter-Efficient Fine-Tuning" https://arxiv.org/pdf/2211.15583.pdf
- A. Breden (2020) "Detecting adverse drug reactions from twitter" https://arxiv.org/abs/2005.06634
- Z. Li,(2020) "An effective emotional expression and knowledge-enhanced method for detecting adverse drug reactions," IEEE Access, vol. 8, pp. 87083–87093

- S. Chen, (2019) "Hitsz-icrc: a report for smm4h shared task 2019-automatic classification and extraction of adverse drug reactions in tweets," (SMM4H) pp. 47–51
- Yu Gu (2020) "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing", arxiv.2007.15779
- Scaboro (2023) "Extensive Evaluation of Transformer-based Architectures for Adverse Drug Events Extraction", DOI:2023.110675
- Wang (2023) "Pre-trained Language Models in Biomedical Domain: A Systematic Survey" arXiv:2110.05006
- Gu (2023) Distilling Large Language Models for Biomedical Knowledge Extraction: arXiv:2307.06439
- Gema (2023) Parameter-Efficient Fine-Tuning of LLaMA for the Clinical Domain: arXiv:2307.03042
- Hu (2023) Improving Large Language Models for Clinical Named Entity Recognition via Prompt Engineering arXiv:2303.16416
- Kosprdic (2024) From Zero to Hero: Harnessing Transformers for Biomedical Named Entity Recognition in Zero- and Few-shot Contexts arXiv:2305.04928
- Gurulingappa (2012) Extraction of potential adverse drug events from medical case reports 10.1186/2041-1480-3-15
- Karimi (2015) Cadec: A corpus of adverse drug event annotations doi: 10.1016/j.jbi.2015.03.010. Epub 2015 Mar 27.
- Zolnoori (2019) The PsyTAR dataset: From patients generated narratives to a corpus of adverse drug events and effectiveness of psychiatric medications doi: 10.1016/j.dib.2019.103838.
- Hiba (2023) Fine-Tuning Transformer Models for Adverse Drug Event Identification A Comparative Study arXiv:1901.08746
- HuggingFace (2024 at 04mar), https://huggingface.co/ BERT-large 'bert-large-uncased', SpanBERT 'SpanBERT/spanbert-base-cased', BioLinkBERT 'michiyasunaga/BioLinkBERT-large', BioBERT v1.1 'dmis-lab/biobert-v1.1' and v1.2 'dmis-lab/biobert-base-cased-v1.2', EnDR-BERT 'cimm-kzn/endr-bert', BioClinicalBERT 'emilyalsentzer/Bio-ClinicalBERT', SciBERT 'allenai/scibert-scivocab-uncased'
- Lee (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text miningarXiv:1901.08746
- Zhang (2022) OPT: Open Pre-trained Transformer Language Models arXiv:2205.01068
- Jiang (2023) Mistral 7B arXiv:2310.06825
- Bolton (2022) Stanford: BioMedLM, https://crfm.stanford.edu/2022/12/15/biomedlm.html
- Luo (2022) BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining arXiv:2210.10341
- Zhang (2023) BiomedGPT: A Unified and Generalist Biomedical Generative Pre-trained Transformer for Vision, Language, and Multimodal Tasks arXiv:2305.17100
- Labrak (2024) BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains arXiv:2402.10373
- Collins (2024) https://github.com/collij22/ADEdetectionJC