

# From Rule-Based to DeepSeek R1 – A Robust Comparative Evaluation of Fifty Years of Natural Language Processing (NLP) Models To Identify Inflammatory Bowel Disease Cohorts

**Date:** 06/07/2025

**Authors:** Dr. Matthew Stammers<sup>1,2,3,4</sup>, Dr. Markus Gwiggner<sup>1,4</sup>, Prof. Reza Nouraei<sup>3,4,5</sup>, Prof. Cheryl Metcalf<sup>4,6</sup>, Prof. James Batchelor<sup>3,4</sup>

**Affiliations:**

1. University Hospital Southampton, Southampton, SO16 6YD, United Kingdom
2. Southampton Emerging Therapies and Technologies (SETT) Centre, Southampton, SO16 6YD, United Kingdom
3. Clinical Informatics Research Unit (CIRU), Coxford Road, Southampton, SO16 5AF, United Kingdom
4. University of Southampton, Southampton, SO17 1BJ, United Kingdom, UK
5. Queen's Medical Centre, ENT Department, Nottingham, Nottingham, England, NG7 2UH, United Kingdom, UK
6. University of Southampton, School of Healthcare Enterprise and Innovation, University Of Southampton Science Park, Enterprise Road, Chilworth, SO16 7NS, Southampton, UK

## 1.1 Abstract:

### 1.1.1 Background:

Natural language processing (NLP) can identify cohorts of patients with inflammatory bowel disease (IBD) from free text. However, limited sharing of code, models, and datasets continues to hinder progress, and bias in foundation large language models (LLMs) remains a significant obstacle.

### 1.1.2 Objective:

To evaluate 15 open-source NLP models for identifying IBD cohorts, reporting on document-to-patient-level classification, while exploring explainability, generalisability, bias and cost factors.

### 1.1.3 Design

Fifteen algorithms were assessed, covering fifty years of NLP development: regular expressions, Spacy, bag of words (BOW), term frequency inverse document frequency (TF IDF), Word2Vec, two sentence-based SBERT models, three BERT models (distilBERT, RoBERTa, bioclinicalBERT), and five large language models (LLMs): [Mistral-Instruct-0.3-7B, M42-Health/Llama3-8B, Deepseek-R1-Distill-Qwen-32B, Qwen3-32B, and Deepseek-R1-Distill-Llama-70B]. Models were evaluated based on F1 score, bias, environmental costs (in grams of CO<sub>2</sub> emitted), and explainability.

### 1.1.4 Results:

A total of 9311 labelled documents were evaluated. The fine-tuned DistilBERT model achieved the best performance (F1: 94.06%) and was more efficient (230.1g CO<sub>2</sub>) than all other BERT and LLM models. BOW was also strong (F1: 93.38%) and very low cost (1.63g CO<sub>2</sub>). LLMs performed less well (F1: 86.65% to 91.58%) and had a higher compute cost (938.5 to 33884.4g CO<sub>2</sub>), along with more bias.

### 1.1.5 Conclusion:

Older NLP approaches, such as BOW, can outperform modern LLMs in clinical cohort detection when properly trained. While LLMs do not require task-specific pretraining, they are slower, more costly, and less accurate. All models and weights from this study are released as open source to benefit the research community.

## 1.2 Introduction

### 1.2.1 Key Messages:

- **What is already known about this topic-** *Language models can identify IBD patient cohorts from clinical free-text records, albeit with only moderate accuracy.*
- **What this study adds –** *This study demonstrates that tried-and-tested NLP methodologies can and do generally outperform LLMs at IBD cohort identification. While they do have higher precision than other methods, LLMs encounter significant speed, bias, and cost limitations. This study also reveals faulty assumptions regarding the simplicity of the document-level free text cohort identification task.*
- **How this study might affect research, practice, or policy—***This study highlights significant clinical bias, sustainability, and scalability issues with LLMs, which must be addressed at all levels before practical production clinical deployment can safely occur.*

## 1.2.2 Natural Language Processing in Inflammatory Bowel Disease

Ulcerative colitis (UC), Crohn's disease (CD) and inflammatory bowel disease unclassified (IBD-U) are chronic inflammatory conditions collectively referred to as inflammatory bowel disease (IBD)(1) diagnosed through a combination of clinical, biochemical, genetic, radiological, endoscopic and histopathological tests(2). IBD is known to be less common in non-caucasian populations; however, the incidence of the disease now appears to be increasing(3) among these groups. Natural language processing (NLP) can algorithmically identify patients and perform case matching for clinical trial recruitment(4). However, relatively little robust NLP research has been conducted in gastroenterology to date(5).

Data fragmentation is a known major obstacle(6). Applying NLP algorithms to clinical free text is one of the few ways to address this issue at scale. Rule-based (RB) data trawls using regular expressions (regex) demonstrate high sensitivity but lower precision, with varying overall efficacy across databases (F1: 0.79-0.9). Consequently, better methods are needed to resolve this database fragmentation problem, not only in IBD but also in numerous patient-identification tasks.

Machine learning (ML) NLP algorithms have undergone significant improvements, particularly over the past 50 years. The earliest ML text-classification algorithms took the form of a 'bag-of-words' (BOW) word vector representations(7) developed in 1975. These models, in their simplest form, derive counts of words appearing in a document and associate these counts with a class (during training) to later make classification decisions(8). The original pioneers of ML NLP research subsequently devised other methods, such as Term Frequency Inverse Document Frequency (TF-IDF)(9), a form of vector space model(10) focused on rarer words and other similar, simpler NLP document classification models(11).

However, the primary limitation of these classification algorithms was their inability to understand context or complex associations between words. Consequently, despite considerable work between 1990 and 2017, progress in NLP was only incremental. This, however, changed in 2017 with the advent of the transformer architecture(12). The discovery of the attention head finally enabled computers to focus on written context and begin to understand human language better. Within a year of that paper, bidirectional encoder representations from transformers (BERT)(13) and pre-trained generative transformers (GPT)(14) arrived.

DistilBERT(15) is a lighter and faster version of the original BERT model, operating 60% faster while preserving 95% of BERT's language understanding benchmark performance. In contrast, RoBERTa(16) was trained on over 160GB of uncompressed text. However, while neither of these models was explicitly trained for clinical tasks, BioClinicalBERT(17)

was. In contrast, new open-source GPT models are now released weekly and have garnered significantly more public attention since GPT-3(18) and the public release of ChatGPT™ in 2022. Such large language models (LLMs) perform well on closed benchmarks (MedQA, etc.), but their performance on open medical benchmarks is less impressive despite specialist prompting techniques(19).

In this study, we evaluate five of the current open-source frontrunners(20): M42-Health/Llama3-8B(21), Deepseek-R1-Distill-Qwen2.5-32B(22), Qwen3-32B(23), and Deepseek-R1-Distill-Llama3-70B(22) to assess their zero-shot performance against this novel clinical cohort identification task.

### **1.2.3 Aim**

This study develops and thoroughly validates open-source document classification models for IBD, exploring the concepts of explainability, cost, and bias in depth.

### **1.2.4 Objectives**

1. Develop, validate, and prioritise methods based on RB, ML, and foundation models (LLMs) for identifying IBD patients.
2. Identify biases, economic impacts, and other costs associated with model inference.
3. Investigate interactions in document-level and patient-level cohort identification as well as model explainability.

## 1.3 Methods

### 1.3.1 Inclusion Criteria

All adults aged 18 and over who were first electively referred to the tertiary academic teaching hospital for specialist gastroenterology care between 2007 and 2023, and who did not opt out of allowing their clinical data to be used for secondary care research, were considered for inclusion in the study.

### 1.3.2 Checklist and Ethics

The study adheres to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) for AI checklist(24). Full details of this checklist are provided in *Error! Reference source not found.*. The study protocol was submitted to the Wessex REC (23/SC/0152), which provided research ethics board approval. The study was registered locally as RHM MED1947 on 22/03/2023.

### 1.3.3 Primary & Secondary Outcomes

- Primary outcomes of interest were the harmonic F1 score ( $2 * \text{precision (PPV)} * \text{recall (sensitivity)} / \text{precision} + \text{recall}$ ) and Matthews correlation coefficient (MCC) for IBD diagnosis at both patient and document level.
- Secondary outcomes of interest were fairness statistics(DI/EO/DP), time (s), energy (kWh), CO2 production (grams), model Brier scores and Gini coefficients.

### 1.3.4 Study Configuration

#### 1.3.4.1 Data Sources

This study focuses on gastroenterology letters, endoscopy reports, and histopathology reports.

#### 1.3.4.2 Data Pre-Processing and Quality Checking

All data were handled using a SQL pipeline per NHS reproducible analytics pipelines (RAP) best practices(25) using a robust test-driven development (TDD) approach without filtering other than free-text clinical data redaction using Pteredactyl(26), a locally developed Python-based free-text redaction wrapper around Presidio(27), employing the Stanford base de-identifier model(28) to mask patient-identifiable information (PII). The text was then manually screened to remove any PII that Pteredactyl missed(29).

#### **1.3.4.3 Gold Standard Cohort Identification, Data Linkage & Outcome Labelling**

A team of three junior doctors, led by a gastroenterology registrar, initially conducted partially blinded manual chart reviews on a randomly selected cohort of suspected IBD patients comprising 2,800 individuals(6). Then, a subset of this cohort, comprising 1,612 patients, was identified, with available and linkable free-text documents.

Free-text documents were chronologically linked, starting with endoscopy records matched to histopathology reports if the procedure occurred within 72 hours before sample receipt, and the histological type aligned. Relevant preceding or following clinic letters were then added. A consultant gastroenterologist (MS, 14 years' experience) re-validated all linked records, averaging 5.78 documents per patient. A strict IBD definition was applied, with any diagnostic ambiguity (e.g., 'possible' or 'potential' IBD) deemed non-diagnostic to maximise classifier precision. Because the primary purpose of the algorithm was to identify all patients who should be under our local IBD service, patients with microscopic colitis were included in the capture, but all other forms of colitis were excluded. Microscopic colitis is strongly associated with IBD(30). At an IBD service planning level, it is appropriately included. However, there may be a demand for these models to be used at a population level in future. Re-training the models to achieve this would be a relatively trivial matter.

#### **1.3.4.4 Predictor Handling**

The training and validation sets for the trained models were divided 70/30 (at a patient level), and each model's 30% holdout set was utilised exclusively for validation with checks in place to prevent data leakage. However, LLMs were evaluated against the entire set because this study for them represents type IV validation(31) because there is no way these LLMs could have ever seen the source data before.

#### **1.3.4.5 Platform Hardware, Software & Dev-Ops**

The platform itself was set up as described in *Error! Reference source not found.*. Here, you can find all the links to the code, along with detailed information about the physical hardware setup.

#### **1.3.4.6 LLM Prompt Templating**

A JSON-based zero-shot query method was employed to assess the LLMs, and this process is described in more detail in *Error! Reference source not found.*. This template enabled attempts to be made to assess LLM calibration, and by building on MedPrompt(19), leveraging the Clue and Reasoning Prompt (CARP) method(32), facilitating state-of-the-art document classification performance.

### 1.3.5 Analytical Methods

#### 1.3.5.1 Sample Size Calculation

The outcome of this study is a binary classification between 'IBD' and 'Not-IBD'. The most significant factors influencing sample size are class imbalances in the training and validation cohorts, which were nearly 8 to 2 IBD to Not-IBD(6).

Using Pate & Riley's formula(33) with a predicted  $r^2$  of 0.05, shrinkage of 0.9, a prevalence of 0.8 and 4 candidate predictors (4 data streams with clinic letters preceding and following the endoscopies/histopath are analysed separately), a sample size of 542 is suggested as required to power the study sufficiently. However, the formula above likely underestimates the actual number of candidate predictors, which cannot be known at this point. Juckett et al.'s (34) work suggests that rare tokens carry less predictive weight and that once samples exceed 1,000 records, a capture probability of > 95% is typically attained. If all 4 document types were vastly different, then a minimum sample size of 4,000 would be required. Because a consultant(attending) physician was performing the study, bringing down the cost of labelling, in the end, nearly 10,000 documents were annotated to absolutely guarantee sufficient power for the study.

#### 1.3.5.2 Study Metrics

A complete set of measurement metrics is used in this study, as highlighted in *Error! Reference source not found.*

#### 1.3.5.3 Statistical Analyses

Means and medians were calculated as appropriate based on skewness, using a 95% confidence interval or 25%/75% quantiles. Model performance was compared by age, sex, and the index of multiple deprivations decile (IMD), with 10 being the least deprived. Statistical tests were performed as appropriate, with Chi2, Pearson correlation, and Mann-Whitney U tests being used predominantly due to the non-parametric nature of the data. 'Wilson's confidence intervals for binomial proportions were calculated using the stats module from Scipy, which provides better coverage probabilities than a normal approximation.

#### 1.3.5.4 Document/Patient-Level Interactions

Decision tree (DT) algorithms were used to determine the optimal fit between document-level and patient-level predictions. Gini coefficients assessed the purity of each branch in the tree's logic, with a value of 0, indicating perfect separation. Logistic Regression (LR) classifiers were used in many of the NLP pipelines, but these are described in more detail below in *Model Setup*.

### 1.3.5.5 Handling Class Imbalance

Due to class imbalance, the harmonic F1 score was preferred as the primary metric for outcome measurement, followed by the Matthews Correlation Coefficient (MCC).

### 1.3.5.6 Cross Validation & Calibration

Discrimination was assessed using the Brier score(35) and model calibration was evaluated visually using calibration plots. Cross-validation was performed as per **Predictor Handling**. Feature selection was performed by the models as described in **Model Setup**, with the final designs decided upon after much experimentation and records of the older experiments retained.

### 1.3.5.7 Missingness Transformations & Protecting Clinical Free-Text Information

The text was fully pseudonymised as described in **Data Pre-Processing and Quality Checking**. Protecting PII was prioritised over any negative impacts on algorithm performance caused by masking. Missingness was quantified and emphasised where relevant. No other transformations were applied to the data as it did not require winsorising or scaling.

### 1.3.5.8 Error Analysis

Any significant outliers or abnormal results were re-examined through carefully stored outputs in folders throughout the pipeline, along with tests to manage the system. The code was fully documented with docstrings and complete tests to minimise errors. If an issue is discovered, please raise it as an issue on the [GitHub page](#).

### 1.3.5.9 Fairness/Bias Evaluation

Fairness evaluation was conducted on binned demographic characteristics for every model using demographic parity (DP)(36), equal opportunity (EO)(37) and disparate impact (DI)(38) statistics. All fairness analyses were performed on the validation set only.

### 1.3.5.10 Economic & Sustainability Analysis

Inference time and computation costs were calculated for each model in succession(39). Emission factors are derived from UK government statistics and the 2024 conversion factors(40), which are set at 0.20705 kg CO<sub>2</sub>e per kWh according to the 2024 guidance to provide a level playing field for comparison. Calculating the precise energy usage and carbon footprint of LLMs is more challenging because they typically operate across multiple GPUs simultaneously, with tests often taking several days to complete through a local network API, which can be more easily interrupted. Therefore, the best estimates were derived using average watt consumption per hour and algorithm runtime.

### 1.3.5.11 Explainability Analysis

To better understand model predictions, SHapely Additive exPlanations (SHAP)(41,42)-2014 and Local Interpretable Model-agnostic Explanations(LIME)(43)-2016 were both used. These are the two most popular ML explainability methods presently available(44).

### 1.3.6 Model Setup

Fifteen models are analysed in this study, falling into three primary groups. The first five models [regex, spacy, bag-of-words (BOW)(7), term frequency-inverse document frequency (TF-IDF)(9) and word-2-vector (word2vec)] models are all trained from scratch. The next five transformer-based models are all fine-tuned: [sBERT(45), sBERT-med(46), distilBERT(15), bio-clinicalBERT(15)[BioClinicalBERT] and RoBERTa(47)] and the final five, all GPT-based models were managed solely via prompt engineering [Mistral-Instruct-0.3-7B(20), M42-Health/Llama3-8B(21), Deepseek-R1-Distill-Qwen2.5-32B(22), Qwen3-32B(23), and Deepseek-R1-Distill-Llama3-70B(22)].

Part of the reason for publishing the code fully open source is to allow other developers and data scientists to inspect for themselves how each algorithm was handled. However, for convenience, a complete description of the handling of each model is provided in *Error! Reference source not found.* to make this explicit for all readers.

## 1.4 Results

### 1.4.1 Total Study Cohort

From the randomly selected gold-standard cohort of 2,800 patients, 1,612 individual patients were found to have chronologically linkable endoscopic and histopathological records, as per **Figure 1**. Eighty-nine patients did not have chronologically linkable clinic letters available.

#### **Figure 1: Study Population CONSORT**

Figure\_1

*Figure 1 describes the study population in a CONSORT diagram*

#### 1.4.1.1 Training and Validation Cohorts

The training set contained 1,128(70%) patients, and the validation set contained 484(30%). Within the training set, 872(77%) of patients had IBD, and 258(23%) did not. Within the validation set, 378(78%) of patients had IBD, and 110(22%) did not.

In total, 9,311 free-text documents were manually reviewed. The training set contained 6,559 documents, of which 4492(68%) were labelled as suggestive of IBD and 2097(32%) were not. The validation set contained 2,752 free-text documents, of which 1796(65%) were labelled as suggestive of IBD and 956(35%) were not. There were 2,592 rows of carefully aligned document data – 1824 in the training set and 768 in the validation(test) set. Coverage is reported according to the number of these rows used. There were no significant differences between the development and validation settings, eligibility criteria, outcome and predictors.

#### **1.4.1.2 Cohort Demographics**

No significant differences were observed between the training and validation cohorts in any continuous demographic. However, the validation cohort had ~3% more females, and the validation cohort was ~2% less white. These were the only two significant results. Full demographic results are provided in *Error! Reference source not found.*

#### **1.4.2 Predictive modelling**

##### **1.4.2.1 UMLS vs Free Text**

The addition of UMLS had a detrimental effect on overall performance. This is because the meta-thesaurus, even though it was carefully filtered, still captured some terms inappropriately connected to IBD, such as ‘17-hydroxycorticosterone’ and ‘vinblastine/methotrexate’, which all map to IBD-associated concept unique identifiers (CUIs) within UMLS. The PTs are thus vulnerable to over-mapping across ontologies. Accordingly, only the free-text results are reported in full here.

##### **1.4.2.2 LLM Response Quality by Model**

m42 provided the most incorrectly formatted .json responses n=391, followed by mistral, n=5. The larger LLMs rarely produced incorrectly formatted .json responses (n=2).

##### **1.4.2.3 Document Level Full Results**

At a document level, model performance was very variable. Top performers included BOW (F1: 93.84%), SBERT (F1: 94.25%) and the full BERT models (F1: 94.47-95.13%). Full results are given in *Error! Reference source not found.* Most models lack specificity, indicating that they are not very effective at correctly identifying true negatives.

Full development and validation results for each model are provided in the [GitHub repo](#) for complete transparency.

### 1.4.2.4 Patient-Level Full Results With Patient-Level Trained Models

At a patient level, model performance degrades by approximately 2% of the F1 score if the models are not retrained and 1% if they are (**Table 1**). BOW and the ‘all-mpnet-base-v2’ SBERT model performed exceptionally well. Interestingly, the medically finetuned SBERT model performed comparatively poorly and RoBERTa suffered a catastrophic collapse during the patient-level prediction task.

**Table 1: Patient Level Prediction Results**

Model	Doc Coverage	Accuracy	Precision	Recall	Specificity	NPV	F1 Score	MCC
<b>Fully Trained</b>								
<b>Regex</b>	768 (100.00%)	81.51% (CI: 78.61% - 84.10%)	81.24% (CI: 78.31% - 83.86%)	100.00% (CI: 99.38% - 100.00%)	7.19% (CI: 4.06% - 12.41%)	100.00% (CI: 74.12% - 100.00%)	89.65% (CI: 89.26% - 90.04%)	0.2417 (CI: 0.1623 - 0.3016)
<b>Spacy</b>	768 (100.00%)	80.99% (CI: 77.26% - 84.24%)	80.34% (CI: 76.50% - 83.69%)	100.00% (CI: 98.99% - 100.00%)	14.81% (CI: 9.33% - 22.72%)	100.00% (CI: 80.64% - 100.00%)	89.10% (CI: 88.37% - 89.84%)	0.3450 (CI: 0.2568 - 0.4168)
<b>BOW</b>	768 (100.00%)	88.84% (CI: 85.73% - 91.35%)	89.02% (CI: 85.70% - 91.64%)	98.20% (CI: 96.32% - 99.12%)	51.04% (CI: 41.20% - 60.81%)	87.50% (CI: 76.37% - 93.81%)	93.38% (CI: 92.16% - 94.59%)	0.6138 (CI: 0.5201 - 0.7015)
<b>TF-IDF</b>	768 (100.00%)	83.68% (CI: 80.12% - 86.70%)	83.37% (CI: 79.71% - 86.48%)	99.48% (CI: 98.14% - 99.86%)	19.79% (CI: 13.05% - 28.86%)	90.48% (CI: 71.09% - 97.35%)	90.72% (CI: 89.86% - 91.69%)	0.3773 (CI: 0.2729 - 0.4738)
<b>Word2Vec</b>	768 (100.00%)	81.82% (CI: 78.14% - 85.00%)	81.51% (CI: 77.78% - 84.74%)	100.00% (CI: 99.02% - 100.00%)	8.33% (CI: 4.28% - 15.59%)	100.00% (CI: 67.56% - 100.00%)	89.81% (CI: 89.30% - 90.44%)	0.2606 (CI: 0.1588 - 0.3340)
<b>Finetuned</b>								
<b>SBERT</b>	768 (100.00%)	89.67% (CI: 86.64% - 92.08%)	88.81% (CI: 85.44% - 91.48%)	99.20% (CI: 97.68% - 99.73%)	56.48% (CI: 47.07% - 65.45%)	95.31% (CI: 87.10% - 98.39%)	93.72% (CI: 92.50% - 94.90%)	0.6844 (CI: 0.6120 - 0.7545)
<b>SBERT Med</b>	768 (100.00%)	84.71% (CI: 81.23% - 87.64%)	84.58% (CI: 80.97% - 87.61%)	98.97% (CI: 97.38% - 99.60%)	27.08% (CI: 19.20% - 36.73%)	86.67% (CI: 70.32% - 94.69%)	91.21% (CI: 90.07% - 92.38%)	0.4308 (CI: 0.3263 - 0.5233)
<b>DistilBERT</b>	768 (100.00%)	90.08% (CI: 87.10% - 92.44%)	90.48% (CI: 87.29% - 92.93%)	97.94% (CI: 95.98% - 98.95%)	58.33% (CI: 48.34% - 67.69%)	87.50% (CI: 77.23% - 93.53%)	94.06% (CI: 92.59% - 95.30%)	0.6624 (CI: 0.5724 - 0.7458)
<b>BioclinicalBERT</b>	768 (100.00%)	90.29% (CI: 87.33% - 92.62%)	91.48% (CI: 88.39% - 93.81%)	96.91% (CI: 94.67% - 98.22%)	63.54% (CI: 53.57% - 72.48%)	83.56% (CI: 73.43% - 90.34%)	94.12% (CI: 92.79% - 95.48%)	0.6735 (CI: 0.5892 - 0.7538)
<b>RoBERTa</b>	768 (100.00%)	80.17% (CI: 76.38% - 83.47%)	80.17% (CI: 76.38% - 83.47%)	100.00% (CI: 99.02% - 100.00%)	0.00% (CI: 0.00% - 3.85%)	nan% (CI: N/A)	88.99% (CI: 88.99% - 88.99%)	0.0000 (CI: 0.0000 - 0.0000)

Prompt Engineered								
<b>Mistral-0.3-7B</b>	2510 (100.00%)	81.05% (CI: 79.48% - 82.53%)	94.47% (CI: 93.30% - 95.44%)	81.37% (CI: 79.63% - 83.00%)	79.67% (CI: 75.85% - 83.02%)	50.07% (CI: 46.54% - 53.59%)	87.43% (CI: 86.34% - 88.56%)	0.5214 (CI: 0.4850 - 0.5571)
<b>M42-Llama_8B</b>	2510 (100.00%)	81.67% (CI: 80.11% - 83.14%)	93.26% (CI: 92.01% - 94.33%)	83.32% (CI: 81.63% - 84.88%)	74.79% (CI: 70.74% - 78.46%)	51.71% (CI: 48.01% - 55.40%)	88.01% (CI: 86.91% - 89.10%)	0.5112 (CI: 0.4755 - 0.5486)
<b>DeepSee k-R1-Qwen2.5_32B</b>	2510 (100.00%)	80.31% (CI: 78.72% - 81.80%)	95.83% (CI: 94.78% - 96.68%)	79.08% (CI: 77.27% - 80.78%)	85.48% (CI: 82.08% - 88.33%)	49.18% (CI: 45.83% - 52.53%)	86.65% (CI: 85.54% - 87.73%)	0.5391 (CI: 0.5084 - 0.5672)
<b>Qwen3_32B</b>	2510 (100.00%)	80.22% (CI: 78.63% - 81.72%)	96.35% (CI: 95.34% - 97.15%)	78.50% (CI: 76.67% - 80.22%)	87.47% (CI: 84.24% - 90.12%)	49.13% (CI: 45.82% - 52.46%)	86.51% (CI: 85.27% - 87.67%)	0.5478 (CI: 0.5178 - 0.5774)
<b>DeepSee k-R1-Llama70B</b>	1326 (52.82%)	86.95% (CI: 84.87% - 88.77%)	96.68% (CI: 95.25% - 97.69%)	86.99% (CI: 84.69% - 89.00%)	86.73% (CI: 81.49% - 90.66%)	60.00% (CI: 54.41% - 65.34%)	91.58% (CI: 90.25% - 92.89%)	0.6464 (CI: 0.5979 - 0.6938)

Table 1 Describes the full patient-level analysis results for the various NLP models in the study.

LLMs were generally better at patient-level prediction but struggled with document-level prediction vs other models, suggesting that they perform better with greater available context.

#### 1.4.2.5 Model Calibration

Brier scores ranged from 0.711 (DistilBERT) to 0.121 (SBERT Med) for the document classification task and from 0.093 (DistilBERT) to 0.1983 (RoBERTa) on the patient-level prediction task. RoBERTa was trained identically to the other two BERT models with the same hyperparameters. Its catastrophic collapse is purely due to differential pre-training and tokenisation, leading it to predict only positives correctly at a patient level. Such is the potentially brittle nature of these algorithms. Clinical pre-training did not cause the bioclinicalBERT model to outperform distilBERT. The same was true with the SBERT models, where the base SBERT algorithm was substantially better calibrated (0.0829-0.1178) than the medically pre-trained one (0.1061-0.1954). BOW also performed remarkably well (0.098-0.124) vs other ML algorithms like word2vec (0.1211-0.1901).

LLMs, despite not having been fine-tuned to the task, performed remarkably well. However, their Brier scores ranged from 0.1305 (Deepseek Llama 70B) to 0.1978 (Deepseek Qwen 32B), suggesting that their predictions were comparatively poorly calibrated. This will be assessed in more detail in a subsequent study.

### 1.4.3 Document and Patient-Level Interactions

Notably, performance in the document classification task does not reliably predict performance at the patient-level identification task.

#### 1.4.3.1 Document and Patient-Level Database Correlations

Correlations between individual document types and diagnosis vary dramatically even by database. These effects are highlighted in **Figure 2**, with clinic letters more strongly correlated (0.57-58) with patients ultimately having IBD than endoscopy reports (0.45) or histopathology reports (0.44).

**Figure 2: Pearson Correlations Between Document and Patient-Level IBD Diagnosis**

Figure\_2

*Figure 2 describes the Pearson correlations between gold-standard cohort elements*

This complex, unexpected complication has not even been mentioned in the literature up until this point(5,6,48). These complexities explain a lot of the performance degradation seen when moving from document to patient-level prediction.

#### 1.4.3.2 Document to Patient Regression and Tree Models

A simple L1 (lasso) LR model was used to assess the individual comparative predictive performance of each variable towards a correct diagnosis of IBD. The following clinic letter was the most useful, with odds of 32.89 vs 20.49 for the preceding letters and 5.9-8.65 for endoscopy and histopathology records (**Figure 3**).

**Figure 3: Odds Ratios by Document Type**

Figure\_3

*Figure 33 above highlights the different weights of positive document identification as a contributor to patient-level IBD prediction.*

Finally, a decision tree classifier was developed to visualise and attempt to manage the above matrix if possible. However, the gini coefficient never reached zero at any branching step of the logic tree, even if the following clinic split represented the first branch, suggesting that document-to-patient level mapping is not simply solved.

#### 1.4.3.3 Document to Patient Model Performance Variation

Clinic letters appeared to be the most difficult to classify Preceding: (MCC: 0.279-0.678), following: (MCC:0.173-0.779) with endoscopy being the easiest (MCC: 0.611-0.776) and histopathology being only slightly more difficult (MCC: 0.543-0.763). This suggests that the document models tend to perform best for document types that are less important in

predicting whether individual patients have IBD overall. Although individual models excelled at specific document types, only distilBERT and bioclinicalBERT sustained MCCs above 0.6 across the options presented, suggesting a significant degree of brittleness among the models tested in this experiment.

#### 1.4.4 Fairness Analysis

In *Error! Reference source not found.* the full fairness results are given for each model at a patient level. Overall, LLMs were the most biased, followed by the BERT models. The LLM biases, however, were often opposite to the BERT models, which typically encoded the biases inherent in the fine-tuning population. Rule-based methods were consistently the least biased, and some of the LLMs were very biased, in particular those based on Qwen v 2.5 and 3.

#### 1.4.5 Economic Comparison

A complete economic analysis was undertaken by model, as per **Table** .

**Table 2: Full Economic Results By Model**

<b>Model</b>	<b>Training + Inference Time (minutes)</b>	<b>Total kWh</b>	<b>CO2 Emissions (grams)</b>
<i>Regex</i>	1.56	0.005	1.04
<i>Spacy</i>	12.48	0.030	6.21
<i>BOW</i>	1.46	0.00786	1.63
<i>TF-IDF</i>	1.48	0.0079	1.64
<i>Word2Vec</i>	33.65	0.18	37.36
<i>SBERT</i>	5.61	0.045	9.31
<i>SBERT Med</i>	2.68	0.019	3.96
<i>DistilBERT</i>	119.71	1.11	230.09
<i>BioclinicalBERT</i>	213.55	2.01	416.73
<i>RoBERTa</i>	228.29	2.13	441.75
<i>Mistral-0.3-7B*</i>	1075	4.54	938.50
<i>M42-Llama_8B*</i>	1080	4.55	942.90
<i>DeepSeek-R1-Qwen2.5_32B*</i>	4142	27.75	5,745.90
<i>Qwen3_32B*</i>	4146	27.78	5,751.50
<i>DeepSeek-R1-</i>	18050	163.65	33,884.40

## Llama70B\*

Table describes the full economic and sustainability analysis results for the included models. [\* No Training - Inference Only.]

Fine-tuning BERT models is moderately costly in terms of computation time, but inference is comparatively rapid. SBERT models are lightweight and can be faster than even running Spacy phrasematcher pipelines, yielding better results. Regex, BOW and TF-IDF take the crown for speed and efficiency at this task. Running inference with LLMs is slow, and, in the case of the largest models, it is prohibitively costly, particularly when the degree of associated performance gain is considered.

### 1.4.6 Explainability Analysis

The string-based methods are entirely explainable. SHAP plots for BOW, TF-IDF, and Word2Vec all highlight similar word token patterns, as highlighted in *Error! Reference source not found.* SHAP plots for distilBERT suggest that the tokens 'microscopic', 'mapping', and 'biopsy' were positively weighted in the model, while bioclinicalBERT emphasised 'mapping', 'reporting', and 'terminal'. RoBERTa, however, weighted 'specialist', 'colon', 'pain', 'inflammatory', and 'external', indicating that this model developed a different understanding of the data than the other two BERT models. These features are prevalent in the dataset, likely contributing to the model's collapse, as it predicts every instance as IBD. LIME analysis of the same models assigned negative weights to terms like 'bowel' and 'biopsies'. Examining one of these examples visually, as per *Error! Reference source not found.*, proves to be illuminating as it demonstrates how sensitive transformers are to self-contradictory information in a clinical record.

LLM performance was overall much the same, with contradictory information and statements, such as 'possible' IBD, which often confused them. In several instances, the LLMs did not realise that microscopic colitides were forms of IBD. This aspect of LLM explainability will be delved into in much more detail in a future study.

## 1.5 Discussion

This study has highlighted numerous challenges in identifying clinical IBD patient cohort groups through NLP, many of which have not been previously documented(5,6,48). The study emphasises that patient-level cohort identification is somewhat context-dependent and cannot be conducted in isolation. BOW and some types of BERT models (SBERT, distilBERT, and bioclinicalBERT) proved to be the most effective; however, other algorithms, like LLMs, demonstrate useful emergent qualities even if fairness and cost limit their scale. This example highlights, in a microcosm, one of the considerable problems with foundation models and why they cannot be used safely in a healthcare

context without further development work. In particular, Qwen 2.5 and 3 are shown to be substantially biased. Simple rule-based methods are by far the most effective at maximising recall and have the added advantages of low bias and cost while remaining fully interpretable.

It should be simple for others to implement the models in other English-speaking contexts now that they are released open source, because very limited pre-processing is required. A simple set of [demo apps](#) is made available for users to demonstrate how easy the models are to use. The models should function correctly whether PII is redacted or not, and they should operate with either the transformers or scikit-learn libraries without difficulty. The LLM framework is proven to generalise in this study; however, the generalisability of other models remains to be established.

Fairness is an issue for all types of models, except for rule-based classifiers. Locally trained models reflect the biases of the local population, while LLMs typically contain opposing but generally more extreme biases. Model collapse was an issue with some models, such as RoBERTa, suggesting that pre-training size alone is not particularly beneficial. Weaknesses of this study include class imbalance within the high-yield, high-information-availability training and validation cohort. Many of these issues are mitigated through the use of the F1 score and MCC as primary evaluation metrics. The sample size, methodology, and low levels of missingness are strong features of this study. Another potential criticism of the study is the use of 8-70B parameter LLMs due to hardware restrictions. While there is a slight chance that the 'full' R1, Mistral, or Qwen model would have performed significantly better, the evidence we have suggests that large LLMs are, in fact, even less faithful when managing factual information, even if readability and informativeness improve(49). A subsequent study is planned to inspect the quality of LLMs at this task.

The strengths of this study include the level of detail provided in the analysis, the transparent reporting methods employed, and the open sharing of source code and models, which is essential for substantial progress being made in this field(5). The validation of the gold standard cohort in this study was extremely robust and led by a senior gastroenterologist with strong informatics experience. Unfortunately, since the data was not consented to, it cannot be shared at present. One of the main weaknesses of this study was the inclusion of patients with microscopic colitis in the IBD cohort. Technically speaking, microscopic colitis is not IBD, even if it might be a precursor(50). This was added for local service reasons. However, if there is strong demand for it, then the models will be retrained without this inclusion – a relatively simple undertaking. This highlights another major weakness of this study, namely its single-site nature. Conducting multi-site studies like this isn't easy due to the localised nature of high-performance computing

(HPC) infrastructure. Hopefully, with the arrival of the secure data environments (SDEs) this will soon change(51).

Other authors do not seem to be aware of the document-to-patient mapping problem(52–57), with most studies primarily only focusing on extracting clinical concepts from annotated free-text clinical records. Future planned work will explore optimal methods to enhance LLM performance in cohort identification and document classification tasks and assess hybrid model-based methods to identify clinical cohorts successfully.

## 1.6 Conclusion

Older NLP technologies from over 50 years ago, such as BOW, can outperform even the most modern NLP technologies, such as Deepseek R1, when properly trained to detect clinical cohorts. Although they don't require pre-training, LLMs are cumbersome, slow and expensive to use for cohort identification tasks and tend towards low recall.

All the models trained in this study are made [open source/weight](#) to benefit patients and the broader research community.

### 1.6.1 Published Declarations Alongside Manuscript

**Word Count:** 3,998

**Twitter:** Matt Stammers: @MattStammers\_

**Email:** m.stammers@soton.ac.uk

**ORCID:** <https://orcid.org/0000-0003-3850-3116>

**Author Contributors:** MS performed all analyses and prepared the final data. MG, RN, CM, and JB (MS's supervisors) provided critical feedback regarding the manuscript. **MS** is the primary guarantor for the review and the corresponding author.

**Acknowledgements:** The local SETT data and AI, CIRU, and Gastroenterology/IBD teams are acknowledged for building the wider infrastructure that made this project possible.

**Patient consent for publication:** Not applicable

**Data Availability:** Data sharing is not possible in this study because the data were not collected for this purpose; however, secondary data can be made available upon request.

**Code Availability:** All codes used in the analytics for this project are made available open source on GitHub at

[https://github.com/MattStammers/An\\_Open\\_Source\\_Collection\\_Of\\_IBD\\_Cohort\\_Identifier\\_Models](https://github.com/MattStammers/An_Open_Source_Collection_Of_IBD_Cohort_Identifier_Models)

**Model Availability:** Models are made fully accessible open source at:

[https://huggingface.co/collections/MattStammers/a-collection-of-ibd-bert-models-](https://huggingface.co/collections/MattStammers/a-collection-of-ibd-bert-models-682b01badbaa646380f54b14)

[682b01badbaa646380f54b14](https://huggingface.co/collections/MattStammers/a-collection-of-ibd-bert-models-682b01badbaa646380f54b14) and in the [GitHub](#) repo. The LLM's are all open source and can be accessed at the following links:

1. **Mistral 7b-v0.3-Instruct**(58): <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>
2. **Llama3-Med42-8b**(59): <https://huggingface.co/m42-health/Llama3-Med42-8B>
3. **Deepseek-Qwen2.5-32B**(60): <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B>
4. **Qwen-v3-32**(61): <https://huggingface.co/Qwen/Qwen3-32B>
5. **Deepseek-Llama-70B**(62): <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B>

**Patient and Public Involvement:** An IBD patient from our local patient panel contributed to the development of the ethics application and study protocol.

**Human Ethics and Consent to Participate Statement:** The Wessex REC and HRA provided research ethics board approval for this study (IC-IBD -23/SC/0152) on 16/05/2023 (<https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/ic-ibd-ibd-cohort-identification-study/>).

**Competing Interests:** RN received an educational grant from Pentax Medical. MS and MG attended the fully funded Dr. Falk Symposium on AI in Gastroenterology in April 2024.

### **1.6.2 Funding**

This study was indirectly funded by the Southampton Academy of Research (SoAR), which funded some of MS's time as part of UHSFT's Research Leaders Program. Study sponsorship was provided by the UHS Research and Development (R&D) Department. The protocol was independently developed.

## Bibliography

1. Baumgart DC, Sandborn WJ. Inflammatory bowel disease: clinical aspects and established and evolving therapies. *The Lancet*. 2007 May;369(9573):1641–57.
2. Nikolaus S, Schreiber S. Diagnostics of Inflammatory Bowel Disease. *Gastroenterology*. 2007 Nov 1;133(5):1670–89.
3. Barnes EL, Loftus EV, Kappelman MD. Effects of Race and Ethnicity on Diagnosis and Management of Inflammatory Bowel Diseases. *Gastroenterology*. 2021 Feb 1;160(3):677–89.
4. Wornow M, Lozano A, Dash D, Jindal J, Mahaffey KW, Shah NH. Zero-Shot Clinical Trial Patient Matching with LLMs. *NEJM AI*. 2025 Jan;2(1):A1cs2400360.
5. Stammers M, Ramgopal B, Owusu Nimako A, Vyas A, Nouraei R, Metcalf C, et al. A foundation systematic review of natural language processing applied to gastroenterology & hepatology. *BMC Gastroenterol*. 2025 Feb 6;25(1):58.
6. Stammers M, Sartain S, Cummings F, Kipps C, Nouraei R, Gwiggner M, et al. Identification of Cohorts with Inflammatory Bowel Disease Amidst Fragmented Clinical Databases via Machine Learning [Internet]. *Research Square*; 2025 [cited 2025 May 12]. Available from: <https://www.researchsquare.com/article/rs-6298636/v1>
7. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. *Commun ACM*. 1975 Nov 1;18(11):613–20.
8. McCallum A, Nigam K. A comparison of event models for naive bayes text classification. In 1998 [cited 2025 Feb 1]. Available from: <https://www.semanticscholar.org/paper/A-comparison-of-event-models-for-naive-bayes-text-McCallum-Nigam/04ce064505b1635583fa0d9cc07cac7e9ea993cc>
9. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Inf Process Manag*. 1988 Jan 1;24(5):513–23.
10. Turney PD, Pantel P. From Frequency to Meaning: Vector Space Models of Semantics. *J Artif Intell Res*. 2010 Feb 27;37:141–88.
11. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. *J Am Soc Inf Sci*. 1990;41(6):391–407.
12. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: *Advances in Neural Information Processing Systems* [Internet]. Curran Associates, Inc.; 2017 [cited 2023 Aug 25]. Available from: [https://proceedings.neurips.cc/paper\\_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html)
13. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Internet]. *arXiv*; 2019 [cited 2023 Aug 25]. Available from: <http://arxiv.org/abs/1810.04805>
14. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving Language Understanding by Generative Pre-Training.

15. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [Internet]. arXiv; 2020 [cited 2025 Feb 1]. Available from: <http://arxiv.org/abs/1910.01108>
16. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach [Internet]. arXiv; 2019 [cited 2025 Feb 1]. Available from: <http://arxiv.org/abs/1907.11692>
17. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings [Internet]. arXiv; 2019 [cited 2025 Feb 1]. Available from: <http://arxiv.org/abs/1904.03323>
18. Floridi L, Chiriatti M. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds Mach.* 2020 Dec 1;30(4):681–94.
19. Nori H, Lee YT, Zhang S, Carignan D, Edgar R, Fusi N, et al. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine [Internet]. arXiv; 2023 [cited 2024 Sep 22]. Available from: <http://arxiv.org/abs/2311.16452>
20. Jiang AQ, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, et al. Mixtral of Experts [Internet]. arXiv; 2024 [cited 2025 Feb 1]. Available from: <http://arxiv.org/abs/2401.04088>
21. Christophe C, Kanithi PK, Raha T, Khan S, Pimentel MA. Med42-v2: A Suite of Clinical LLMs [Internet]. arXiv; 2024 [cited 2024 Sep 23]. Available from: <http://arxiv.org/abs/2408.06142>
22. DeepSeek-AI, Guo D, Yang D, Zhang H, Song J, Zhang R, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning [Internet]. arXiv; 2025 [cited 2025 Feb 25]. Available from: <http://arxiv.org/abs/2501.12948>
23. Zheng X, Li Y, Chu H, Feng Y, Ma X, Luo J, et al. An Empirical Study of Qwen3 Quantization [Internet]. arXiv; 2025 [cited 2025 May 12]. Available from: <http://arxiv.org/abs/2505.02214>
24. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Calster BV, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024 Apr 16;385:e078378.
25. Shenton CR. NHSE-R - Reproducible Analytical Pipelines (RAP) [Internet]. 2023 [cited 2023 Sep 29]. Available from: <https://nhsengland.github.io/nhs-r-reporting/tutorials/rap.html>
26. GitHub - SETT-Centre-Data-and-AI/Pteredactyl: A python module for redaction of personally identifiable information (PII) in clinical free-text [Internet]. [cited 2024 Dec 31]. Available from: <https://github.com/SETT-Centre-Data-and-AI/Pteredactyl>
27. Kotevski DP, Smee RI, Field M, Nemes YN, Broadley K, Vajdic CM. Evaluation of an automated Presidio anonymisation model for unstructured radiation oncology electronic medical records in an Australian setting. *Int J Med Inf.* 2022 Dec 1;168:104880.
28. StanfordAIML/stanford-deidentifier-base · Hugging Face [Internet]. [cited 2023 Sep 19]. Available from: <https://huggingface.co/StanfordAIML/stanford-deidentifier-base>
29. Pteredactyl PII - a Hugging Face Space by MattStammers [Internet]. [cited 2025 May 26]. Available from: [https://huggingface.co/spaces/MattStammers/pteredactyl\\_PII](https://huggingface.co/spaces/MattStammers/pteredactyl_PII)
30. Khalili H, Burke KE, Roelstraete B, Sachs MC, Olén O, Ludvigsson JF. Microscopic Colitis and Risk of Inflammatory Bowel Disease in a Nationwide Cohort Study. *Gastroenterology.* 2020 May;158(6):1574-1583.e2.

31. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med*. 2015 Jan 6;162(1):55–63.
32. Sun X, Li X, Li J, Wu F, Guo S, Zhang T, et al. Text Classification via Large Language Models [Internet]. *arXiv*; 2023 [cited 2024 Sep 22]. Available from: <http://arxiv.org/abs/2305.08377>
33. Pate A, Riley RD, Collins GS, van Smeden M, Van Calster B, Ensor J, et al. Minimum sample size for developing a multivariable prediction model using multinomial logistic regression. *Stat Methods Med Res*. 2023 Mar;32(3):555–71.
34. Juckett D. A method for determining the number of documents needed for a gold standard corpus. *J Biomed Inform*. 2012 Jun 1;45(3):460–70.
35. Rufibach K. Use of Brier score to assess binary predictions. *J Clin Epidemiol*. 2010 Aug 1;63(8):938–9.
36. Goel N, Yaghini M, Faltings B. Non-Discriminatory Machine Learning through Convex Fairness Criteria.
37. Hardt M, Price E, Price E, Srebro N. Equality of Opportunity in Supervised Learning. In: *Advances in Neural Information Processing Systems* [Internet]. Curran Associates, Inc.; 2016 [cited 2025 May 12]. Available from: [https://proceedings.neurips.cc/paper\\_files/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html)
38. Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In: *Proceedings of the 26th International Conference on World Wide Web* [Internet]. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee; 2017 [cited 2025 May 12]. p. 1171–80. (WWW '17). Available from: <https://doi.org/10.1145/3038912.3052660>
39. CodeCarbon — CodeCarbon 3.0.1 documentation [Internet]. [cited 2025 May 12]. Available from: <https://mlco2.github.io/codecarbon/>
40. GOV.UK [Internet]. 2024 [cited 2025 May 13]. Greenhouse gas reporting: conversion factors 2024. Available from: <https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2024>
41. Mosca E, Szigeti F, Tragianni S, Gallagher D, Groh G. SHAP-Based Explanation Methods: A Review for NLP Interpretability. In: Calzolari N, Huang CR, Kim H, Pustejovsky J, Wanner L, Choi KS, et al., editors. *Proceedings of the 29th International Conference on Computational Linguistics* [Internet]. Gyeongju, Republic of Korea: International Committee on Computational Linguistics; 2022 [cited 2025 Jan 3]. p. 4593–603. Available from: <https://aclanthology.org/2022.coling-1.406/>
42. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst*. 2014 Dec 1;41(3):647–65.
43. Ribeiro MT, Singh S, Guestrin C. ‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier [Internet]. *arXiv*; 2016 [cited 2025 May 12]. Available from: <http://arxiv.org/abs/1602.04938>

44. Sali h AM, Raisi-Estabragh Z, Galazzo IB, Radeva P, Petersen SE, Lekadir K, et al. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Adv Intell Syst.* 2025;7(1):2400304.
45. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [Internet]. arXiv; 2019 [cited 2025 May 13]. Available from: <http://arxiv.org/abs/1908.10084>
46. Deka P, Jurek-Loughrey A, P D. Evidence Extraction to Validate Medical Claims in Fake News Detection. In: Traina A, Wang H, Zhang Y, Siuly S, Zhou R, Chen L, editors. *Health Information Science*. Cham: Springer Nature Switzerland; 2022. p. 3–15.
47. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks [Internet]. arXiv; 2020 [cited 2025 May 13]. Available from: <http://arxiv.org/abs/2004.10964>
48. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc JAMIA.* 2014 Apr;21(2):221–30.
49. Mahapatra J, Garain U. Impact of Model Size on Fine-tuned LLM Performance in Data-to-Text Generation: A State-of-the-Art Investigation [Internet]. arXiv; 2024 [cited 2025 May 14]. Available from: <http://arxiv.org/abs/2407.14088>
50. Tulassay Z, Mihaly E, Herszényi L. Microscopic Colitis: A Challenging Disorder. *Dig Dis.* 2020;38(2):117–21.
51. NHS Transformation Directorate [Internet]. [cited 2023 Sep 22]. Secure Data Environments (SDEs). Available from: <https://transform.england.nhs.uk/key-tools-and-info/data-saves-lives/secure-data-environments/>
52. Stidham RW, Yu D, Zhao X, Bishu S, Rice M, Bourque C, et al. Identifying the Presence, Activity, and Status of Extraintestinal Manifestations of Inflammatory Bowel Disease Using Natural Language Processing of Clinical Notes. *Inflamm Bowel Dis.* 2023 Apr 3;29(4):503–10.
53. Montoto C, Gisbert JP, Guerra I, Plaza R, Pajares Villarroja R, Moreno Almazán L, et al. Evaluation of Natural Language Processing for the Identification of Crohn Disease-Related Variables in Spanish Electronic Health Records: A Validation Study for the PREMONITION-CD Project. *JMIR Med Inform.* 2022 Feb 18;10(2):e30345.
54. Ananthkrishnan A.N., Cai T., Savova G., Cheng S.-C., Chen P., Perez R.G., et al. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing: A novel informatics approach. *Inflamm Bowel Dis.* 2013;19(7):1411–20.
55. Gomollón F, Gisbert JP, Guerra I, Plaza R, Pajares Villarroja R, Moreno Almazán L, et al. Clinical characteristics and prognostic factors for Crohn's disease relapses using natural language processing and machine learning: a pilot study. *Eur J Gastroenterol Hepatol.* 2022 Apr;34(4):389–97.
56. Schmidt L, Ibing S, Borchert F, Hugo J, Marshall AA, Peraza J, et al. Automating Clinical Phenotyping Using Natural Language Processing: An Application for Crohn's Disease [Internet]. medRxiv; 2024 [cited 2025 May 21]. p. 2023.10.16.23297099. Available from: <https://www.medrxiv.org/content/10.1101/2023.10.16.23297099v2>
57. Ibing S, Hugo J, Borchert F, Schmidt L, Benson C, Marshall AA, et al. Electronic Health Records-based identification of newly diagnosed Crohn's Disease cases. *Artif Intell Med.* 2025 Jan 1;159:103032.

58. mistralai/Mistral-7B-Instruct-v0.3 · Hugging Face [Internet]. [cited 2025 May 21]. Available from: <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>
59. m42-health/Llama3-Med42-8B · Hugging Face [Internet]. 2024 [cited 2025 May 21]. Available from: <https://huggingface.co/m42-health/Llama3-Med42-8B>
60. deepseek-ai/DeepSeek-R1-Distill-Qwen-32B · Hugging Face [Internet]. 2025 [cited 2025 May 21]. Available from: <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B>
61. Qwen/Qwen3-32B · Hugging Face [Internet]. 2025 [cited 2025 May 21]. Available from: <https://huggingface.co/Qwen/Qwen3-32B>
62. deepseek-ai/DeepSeek-R1-Distill-Llama-70B · Hugging Face [Internet]. 2025 [cited 2025 May 21]. Available from: <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B>





