BMJ Open Gastroenterology

Robust comparative evaluation of 15 natural language processing algorithms to positively identify patients with inflammatory bowel disease from secondary care records

Matt Stammers [1,2] Markus Gwiggner,[3,4] Reza Nouraei,[2,5] Cheryl Metcalf,[6] James Batchelor[2]

To cite: Stammers M, Gwiggner M, Nouraei R, *et al*. Robust comparative evaluation of 15 natural language processing algorithms to positively identify patients with inflammatory bowel disease from secondary care records. *BMJ Open Gastroenterol* 2025;**12**:e001977. doi:10.1136/bmjgast-2025-001977

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjgast-2025-001977).

Received 6 July 2025
Accepted 26 August 2025

For numbered affiliations see end of article.

Correspondence to
Dr Matt Stammers;
m.stammers@soton.ac.uk

ABSTRACT

Objective Natural language processing (NLP) can identify cohorts of patients with inflammatory bowel disease (IBD) from free text. However, limited sharing of code, models, and data sets continues to hinder progress. The aim of this study was to evaluate multiple open-source NLP models for identifying IBD cohorts, reporting on document-to-patient-level classification, while exploring explainability, generalisability, fairness and cost.

Methods 15 algorithms were assessed, covering all types of NLP spanning over 50 years of NLP development. Rule-based (regular expressions, spaCy with negation), and vector-based (bag-of-words (BoW), term frequency inverse document frequency (TF IDF), word-2-vector), to transformers: (two sentence-based sBERT models, three bidirectional encoder representations from transformers (BERT) models (distilBERT, BioclinicalBERT, RoBERTa), and five large language models (LLMs): (Mistral-Instruct-v0.3-7B, M42-Health/Llama-v3-8B, Deepseek-R1-Distill-Qwen-v2.5-32B, Qwen-v3-32B, and Deepseek-R1-Distill-Llama-v3-70B). Models were comparatively evaluated based on full confusion matrices, time/environmental costs, fairness, and explainability.

Results A total of 9311 labelled documents were evaluated. The fine-tuned DistilBERT_IBD model achieved the best performance overall (micro F1: 93.54%), followed by sBERT-Base (micro F1: 93.05%); however, specificity was an issue for both: (67.80–64.41%) respectively. LLMs performed well, given that they had never seen the training data (micro F1: 86.47–92.20%), but were comparatively slow (18–300 hours) and expensive. Bias was a significant issue for all effective model types.

Conclusion NLP has undergone significant advancements over the last 50 years. LLMs appear likely to solve the problem of re-identifying patients with IBD from clinical free text sources in the future. Once cost, performance and bias issues are addressed, they and their successors are likely to become the primary method of data retrieval for clinical data warehousing.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Language models can identify inflammatory bowel disease (IBD) patient cohorts from clinical free-text records, albeit with only moderate accuracy. The most effective commercial models are not widely available.

WHAT THIS STUDY ADDS

⇒ While well-established natural language processing methods are faster and cheaper than large language models (LLMs), the performance gains are now marginal at best. All the models in this study are provided for free to help facilitate IBD cohort identification even in resource-constrained clinical contexts.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study highlights the value of LLMs for patient cohort identification activities and the increasingly important role they will play in future epidemiological research, clinical data warehousing and case identification.

INTRODUCTION

Natural language processing in inflammatory bowel disease

Ulcerative colitis, Crohn's disease and inflammatory bowel disease (IBD) unclassified are chronic inflammatory conditions collectively referred to as IBD[1] diagnosed through a combination of clinical, biochemical, genetic, radiological, endoscopic and histopathological tests.[2] Data fragmentation is a known major obstacle to the accurate identification of patients with IBD in secondary care.[3] Applying natural language processing (NLP) algorithms to clinical free text is one of the few ways to address this issue at scale.[4 5] The purpose of this study was to develop and

test a variety of algorithms to identify all local patients with IBD and help other clinicians do the same.

Rule-based (RB) searches using regular expressions (regex) or negation strategies (spaCy) demonstrate high sensitivity but lower precision, with varying overall efficacy across databases (F1: 0.79–0.9).[3] Consequently, better methods are needed. 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[6] 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.[7] Term frequency-inverse document frequency (TF-IDF),[8] a form of vector space model[9] focuses on rarer words along with other similar, NLP document classification models.[10] However, these models cannot understand context or complex associations between words.

In 2017, with the advent of the transformer architecture[11] everything changed. Within a year of that paper, bidirectional encoder representations from transformers (BERT)[12] and pretrained generative transformers (GPT)[13] arrived. DistilBERT[14] is a lighter and faster version of the original BERT model, operating 60% faster while preserving 95% of BERT's performance. In contrast, RoBERTa[15] was trained on over 160 GB of uncompressed text. However, while neither of these models was explicitly trained for clinical tasks, BioClinicalBERT[16] was. In contrast, new open-source GPT models are now released weekly and have garnered significantly more public attention since GPT-3[17] 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 has up until now been less impressive despite specialist prompting.[18] In this study, five of the 2025 open-source frontrunners are evaluated: Mistral-Instruct-v0.3-7B,[19] M42-Health/Llama-v3-8B,[20] Deepseek-R1-Distill-Qwen-v2.5-32B,[21] Qwen-v3-32B,[22] and Deepseek-R1-Distill-Llama-v3-70B[21] to assess their zero-shot performance against this novel clinical cohort identification task.

Aim

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

Objectives

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

METHODS

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.

Reporting and ethics

The study adheres to the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) for artificial intelligence (AI) checklist.[23] Full details of this checklist are provided in online supplemental appendix A. The study was registered locally as RHM MED1947 on 22 March 2023.

Primary and secondary outcomes

- The primary outcomes of interest were the core Study Metrics for IBD diagnosis at both patient and document levels.
- Secondary outcomes of interest were fairness statistics—online supplemental appendix D, time (s), energy (kWh), CO2 production (grams), model Brier scores, ORs and Gini coefficients.

Data sources, data preprocessing and using UMLS

This study focuses on gastroenterology letters, endoscopy reports, and histopathology reports. See online supplemental appendix B for details of data handling, quality checking and transformations including the use of the Unified Medical Language System (UMLS).

Gold standard cohort identification, data linkage, predictor handling and validation approach

As already described in the prior study[3] a team of three junior doctors, led by a gastroenterology registrar, initially conducted partially blinded manual chart reviews on a randomly selected, chronologically distributed cohort of suspected patients with IBD comprising 2800 individuals.

A subset of this cohort with available and linkable free-text documents, comprising 1612 patients, was identified. Free-text documents were chronologically linked, starting with endoscopy records matched to histopathology reports where the procedure occurred within 72 hours before sample receipt in the lab, and the histological type aligned exactly.

Relevant clinic letters directly preceding or following were then added. A consultant (Attending) gastroenterologist (MS—14 years' experience) then revalidated all linked records, averaging 5.78 documents per patient.

A strict IBD definition was applied, with any diagnostic ambiguity (eg, 'possible' or 'potential' IBD) deemed non-diagnostic to maximise classifier precision for IBD alone. In V.1 of the experiment, patients with microscopic colitis were left in for service reasons. However, in the final version, they too were removed along with all ischaemic, diverticular, radiation, infective and other

colitides, which always comprised the 'Non-IBD' cohort to maximise task difficulty.

The training and test sets for the trained models were randomly divided 70/30 (at a patient level)—seed 42, and each model's 30% holdout set was used exclusively for testing, with checks in place to prevent data leakage—Type IIa validation according to TRIPOD.[24] LLMs were evaluated against the entire set (with the sole exception of Deepseek-R1-Distill-Llama-v3-70B, which was only validated on the test set and a randomly selected 20% of the remaining training set due to a combination of API glitches, slow speed and poor energy/CO2 efficiency). The LLM validation was Type IV according to TRIPOD.[24] Testing was also performed with a different test set seeded 10 at both a document and patient level to ensure robustness.

Platform hardware, software and LLM prompt templating

The platform and UMLS were set up as described in online supplemental appendix B.

A JavaScript Object Notation (JSON)-based zero-shot query method was employed to assess the LLMs because this is the cleanest way of evaluating their base capabilities. This process is described in more detail in online supplemental appendix C. The template enabled attempts to be made to assess LLM calibration by building on MedPrompt,[18] leveraging the Clue and Reasoning Prompt (CARP) method,[25] facilitating state-of-the-art document classification performance. To test LLM fine-tuning effects, a medically fine-tuned LLM (m42) was included in the battery. Malformed JSON return objects, where possible, were repaired and where not possible, runs were repeated. Zero-shot performance was prioritised because it gives a clear unbiased indication of baseline LLM performance characteristics.

Analytical methods

Sample size calculation

Class imbalances in the training and test cohorts were known from the preceding study.[3] Therefore, rather than relying on Pate *et al*'s formula,[26] which predicted a sample size of only 542, Juckett's[27] work suggests that rare tokens carry less predictive weight and that once samples exceed 1000 records, a capture probability of >95% is typically attained. In the worst-case scenario, a minimum sample size of 4000 would therefore be required. Because a consultant physician was leading the study, the labelling cost was reduced, so nearly 10 000 documents were annotated to guarantee sufficient power for the study.

Study metrics and statistical analysis

A complete set of measurement metrics is used in this study, with a table of metrics of interest available in online supplemental appendix D. However, for less technical readers, the study metrics used in this study are described below with the associated clinical questions that each answers using standard True/False (T/F) and Positive/Negative (P/N) abbreviations:

- Accuracy: Accuracy answers the question: What proportion of all predictions is correct, regardless of disease status?*Calc* : $\frac{(TP+TN)}{(TP+FP+TN+FN)}$
- Precision: Precision (positive predictive value) indicates the trustworthiness of a positive result. *Calc* : $\frac{TP}{(TP+FP)}$
- Negative predictive value (NPV): NPV indicates the trustworthiness of a negative result.*Calc* : $\frac{TN}{(TN+FN)}$
- Recall (sensitivity): Recall answers the question: If a patient has the disease, what is the chance the model will detect it?*Calc* : $\frac{TP}{(TP+FN)}$
- Specificity: Specificity answers the question: If the model says a patient has the disease, how likely is it that they have that particular disease?*Calc* : $\frac{TN}{(TN+FP)}$
- Harmonic micro F1-Score: If one cares about both catching disease (recall) *and* being confident in positives (precision), how good is this test overall?*Calc* : $\frac{2 \times (\text{Precision x Recall})}{(\text{Precision + Recall}).}$
- Matthews' Correlation Coefficient (MCC): MCC assesses all the model's predictions across all classes, positive *and* negative. Although more abstract, it is highly resistant to class imbalance effects. *Calc* : $\frac{(TP \times TN)-(FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$. Given the difficulty of the classification task in this study, any model with a patient-level MCC > 0.6 is considered reasonably good, with <0.3 typically considered poor or random classification.

Model performance was compared by age, gender, ethnicity and the index of multiple deprivations decile, with 10 being the least deprived. The other descriptive statistics used in the study are described in detail in online supplemental appendix D. Due to class imbalance, the MCC was preferred as the primary metric for outcome measurement, followed by the harmonic micro F1-Score.

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 to assess ORs for IBD by document and within many of the NLP pipelines, as described in more detail in online supplemental appendix E. To control for biases which might have arisen by having more than one document per patient, the mappings were repeated with a single row per patient data set.

Cross-validation and calibration

Calibration was assessed using the Brier score[28] and visual plots. Cross-validation was performed as per online supplemental appendix E. Final designs for all models were decided on after much experimentation with full version control. Feature selection, error handling and hyperparameter tuning steps by model are described in detail in online supplemental appendix E.

Fairness/bias evaluation

Fairness evaluation was conducted on binned demographic characteristics for every model using demographic parity (DP),[29] equal opportunity (EO)[30] and disparate impact (DI)[31] statistics. See online supplemental appendix D for complete definitions and the calculations of these statistics, but they are explained below in a non-technical manner for clarity.

DP answers the question: 'Is this model giving absolute positive decisions at equal rates to different groups, regardless of true labels?' The result is provided as an absolute difference where +ve values >0.1 suggest the more privileged group gains, and negative values <−0.1 suggest the reverse. Calc: $DP = P(\hat{Y} = 1 \mid A = \alpha) = P(\hat{Y} = 1 \mid A = \beta)$. Key: P is the probability that $\hat{Y}$ - the predicted outcome is the same between A: the attribute (e.g., gender), which can be either vulnerable (α - female) or not vulnerable (β - male).

EO answers the question: 'Among patients who have the disease, are all demographic groups equally likely to be correctly identified by the model?" If there is no bias, then the value should be close to 0. A value higher than 0.1 suggests bias against the group considered more vulnerable. Calc: $EO = P(\hat{Y} = 1 \mid Y = 1, A = \alpha) = P(\hat{Y} = 1 \mid Y = 1, A = \beta)$. Key: P is the probability that $\hat{Y}$ - the recall of the predicted outcome is the same between A: the attribute (e.g., gender), which can be either vulnerable (α - female) or not vulnerable (β - male).

DI answers the question: 'Is this model proportionally giving positive decisions to different groups at the same rate?' This result is expressed as a ratio. If there is no significant bias, this value should be 0.8–1.25. Calc: $DI = \frac{P(\hat{Y}=1 \mid A=\alpha)}{P(\hat{Y}=1 \mid A=\beta)}$. Key: Divide the protected group's positive prediction rate (α) by that of the most-favoured group (β).

All fairness analyses were performed on the patient-level test set only with the patient-level models.

Economic and sustainability analysis

Inference time and computation costs were calculated for each model in succession.[32] Emission factors are derived from UK government statistics and the 2024 conversion factors,[33] which are set at 0.20705 kg CO2e per kWh according to the 2024 guidance. Calculating the precise energy usage and carbon footprint of LLMs is more challenging; therefore, the best estimates were derived using average watt consumption per hour and algorithm runtime.

Explainability analysis

To better understand model predictions, SHapely Additive exPlanations (SHAP)[34 35] –2014, and Local Interpretable Model-agnostic Explanations (LIME)[36]—2016, were both used. These are the two most popular ML explainability methods presently available.[37] LLMs require an entirely different explainability approach, which will be covered in a subsequent work.

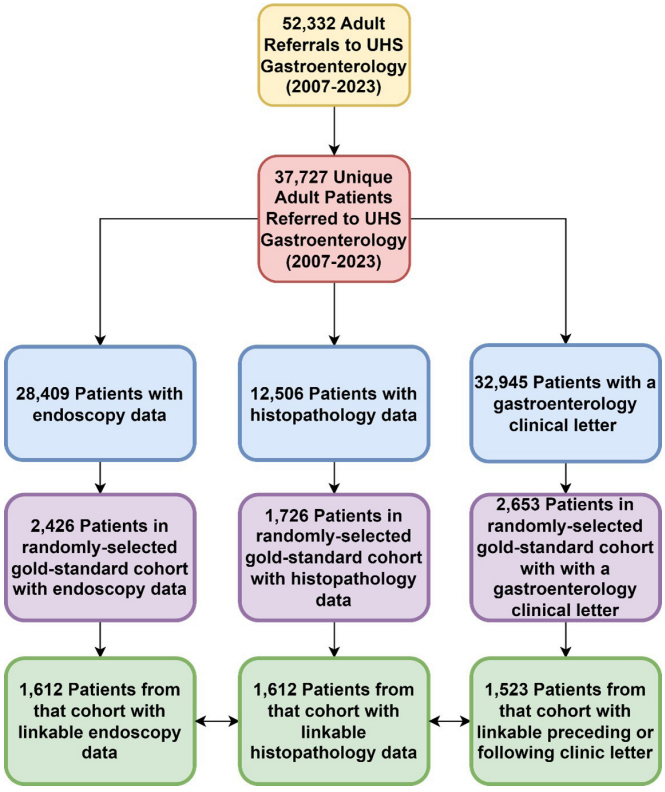

Figure 1 Study population. UHS, University Hospital Southampton.

Model setup

15 models were analysed in this study, falling into three primary groups. The first five models (regex, spaCy, BoW,[6] TF-IDF[8] and word-2-vector (Word2Vec)) models were all trained from scratch. The following five transformer-based models were all fine-tuned: (sBERT,[38] sBERT-med,[39] DistilBERT,[14] BioClinicalBERT[14] and RoBERTa[40]) and the final five, all GPT-based models, were managed solely via prompt engineering (Mistral-Instruct-v0.3-7B,[19] M42-Health/Llama-v3-8B,[20] Deepseek-R1-Distill-Qwen-v2.5-32B,[21] Qwen-v3-32B,[22] and Deepseek-R1-Distill-Llama-v3-70B[21]).

Part of the reason for publishing the code fully open source is to allow other developers and data scientists to improve on the models. For the technically-minded, a complete description of the handling of each model is provided in online supplemental appendix E, along with a full set of references and links.

Patient and public involvement

A patient with IBD from our local patient panel contributed to the development of the ethics application and study protocol.

RESULTS

Total study cohort

Of the 1612 individual patients found to have chronologically linkable endoscopic and histopathological records (figure 1). 89 patients had only linkable endoscopy and histopathology records available.

Table 1 Patient-level prediction results

Model	Doc coverage	Accuracy	Precision	Recall	Specificity	NPV	micro F1-Score	MCC	Brier score
Fully Trained									
Regex	768 (100.00%)	77.48% (CI: 73.55% to 80.98%)	77.05% (CI: 73.06% to 80.61%)	100.00% (CI: 98.96% to 100.00%)	7.63% (CI: 4.06% to 13.86%)	100.00% (CI: 70.09% to 100.00%)	87.04% (CI: 86.52% to 87.66%)	0.2424 (CI: 0.1608 to 0.3150)	0.2252
spaCy	768 (100.00%)	75.83% (CI: 71.82% to 79.43%)	75.00% (CI: 70.88% to 78.71%)	100.00% (CI: 98.92% to 100.00%)	12.03% (CI: 7.54% to 18.65%)	100.00% (CI: 80.64% to 100.00%)	85.71% (CI: 84.99% to 86.46%)	0.3004 (CI: 0.2236 to 0.3629)	0.2417
BoW	768 (100.00%)	87.19% (CI: 83.92% to 89.88%)	88.00% (CI: 84.45% to 90.83%)	96.17% (CI: 93.68% to 97.71%)	59.32% (CI: 50.30% to 67.76%)	83.33% (CI: 73.95% to 89.80%)	91.91% (CI: 90.27% to 93.58%)	**0.6292 (CI: 0.5449 to 0.7057)**	0.1281
TF-IDF	768 (100.00%)	86.98% (CI: 83.69% to 89.69%)	87.41% (CI: 83.82% to 90.29%)	96.72% (CI: 94.36% to 98.11%)	56.78% (CI: 47.77% to 65.36%)	84.81% (CI: 75.30% to 91.09%)	91.83% (CI: 90.36% to 93.42%)	**0.6216 (CI: 0.5353 to 0.7054)**	0.1302
Word2Vec	768 (100.00%)	77.69% (CI: 73.77% to 81.17%)	77.22% (CI: 73.23% to 80.76%)	100.00% (CI: 98.96% to 100.00%)	8.47% (CI: 4.67% to 14.90%)	100.00% (CI: 72.25% to 100.00%)	87.14% (CI: 86.63% to 87.77%)	0.2558 (CI: 0.1799 to 0.3256)	0.2231
Finetuned									
sBERT Base	768 (100.00%)	89.05% (CI: 85.95% to 91.53%)	89.42% (CI: 86.01% to 92.08%)	96.99% (CI: 94.70% to 98.31%)	64.41% (CI: 55.44% to 72.47%)	87.36% (CI: 78.76% to 92.79%)	93.05% (CI: 91.48% to 94.53%)	**0.6866 (CI: 0.6105 to 0.7568)**	0.1095
sBERT Med	768 (100.00%)	80.99% (CI: 77.26% to 84.24%)	80.86% (CI: 76.94% to 84.24%)	98.09% (CI: 96.11% to 99.07%)	27.97% (CI: 20.66% to 36.66%)	82.50% (CI: 68.05% to 91.25%)	88.64% (CI: 87.32% to 89.89%)	0.4063 (CI: 0.3136 to 0.4963)	0.1901
DistilBERT	768 (100.00%)	89.88% (CI: 86.87% to 92.26%)	90.33% (CI: 87.01% to 92.87%)	96.99% (CI: 94.70% to 98.31%)	67.80% (CI: 58.92% to 75.55%)	87.91% (CI: 79.64% to 93.11%)	93.54% (CI: 92.07% to 95.07%)	**0.7120 (CI: 0.6324 to 0.7857)**	0.1012
BioclinicalBERT	768 (100.00%)	89.67% (CI: 86.64% to 92.08%)	90.31% (CI: 86.97% to 92.86%)	96.72% (CI: 94.36% to 98.11%)	67.80% (CI: 58.92% to 75.55%)	86.96% (CI: 78.57% to 92.38%)	93.40% (CI: 91.81% to 94.96%)	**0.7060 (CI: 0.6298 to 0.7795)**	0.1033
RoBERTa	768 (100.00%)	80.99% (CI: 77.26% to 84.24%)	80.31% (CI: 76.39% to 83.71%)	99.18% (CI: 97.62% to 99.72%)	24.58% (CI: 17.69% to 33.06%)	90.62% (CI: 75.78% to 96.76%)	88.75% (CI: 87.68% to 89.88%)	0.4105 (CI: 0.3136 to 0.4976)	0.1901
Prompt Engineered									
Mistral-0.3-7B	2510 (100.00%)	80.22% (CI: 78.63% to 81.72%)	90.57% (CI: 89.12% to 91.84%)	82.72% (CI: 80.98% to 84.34%)	72.12% (CI: 68.40% to 75.56%)	56.32% (CI: 52.79% to 59.79%)	86.47% (CI: 85.34% to 87.65%)	0.5071 (CI: 0.4706 to 0.5435)	0.1895
M42-Llama-8B	2510 (100.00%)	81.79% (CI: 80.23% to 83.25%)	90.11% (CI: 88.65% to 91.40%)	85.44% (CI: 83.78% to 86.95%)	70.22% (CI: 66.44% to 73.73%)	60.29% (CI: 56.61% to 63.84%)	87.71% (CI: 86.63% to 88.70%)	0.5296 (CI: 0.4944 to 0.5656)	0.1833
DeepSeek-R1-Qwen2.5-32B	2510 (100.00%)	83.85% (CI: 82.36% to 85.23%)	95.63% (CI: 94.54% to 96.51%)	82.58% (CI: 80.83% to 84.21%)	87.89% (CI: 85.05% to 90.26%)	61.13% (CI: 57.84% to 64.32%)	88.63% (CI: 87.52% to 89.75%)	**0.6325 (CI: 0.6031 to 0.6598)**	0.1978
Qwen3-32B	2510 (100.00%)	84.53% (CI: 83.08% to 85.88%)	95.54% (CI: 94.45% to 96.42%)	83.62% (CI: 81.91% to 85.19%)	87.48% (CI: 84.61% to 89.88%)	62.47% (CI: 59.17% to 65.66%)	89.18% (CI: 88.21% to 90.19%)	**0.6422 (CI: 0.6127 to 0.6704)**	0.1969
DeepSeek-R1-Llama-70B	1326 (52.82%)	88.12% (CI: 86.24% to 89.78%)	94.03% (CI: 92.35% to 95.36%)	90.43% (CI: 88.44% to 92.11%)	80.14% (CI: 75.14% to 84.34%)	70.77% (CI: 65.60% to 75.45%)	92.20% (CI: 91.02% to 93.38%)	**0.6762** (CI: 0.6272 to 0.7201)	0.1305

Full patient-level analysis results for the various NLP models in the study. The MCC's for the 8 well-performing models are highlighted in bold for clarity.
BOW, bag of words; MCC, Matthews' Correlation Coefficient; NLP, natural language processing; NPV, negative predictive value; regex, regular expressions; SBERT, sentence-bidirectional encoder representations from transformers; Spacy, negation strategies; TF IDF, term frequency-inverse document frequency; Word2Vec, word-2-vector.

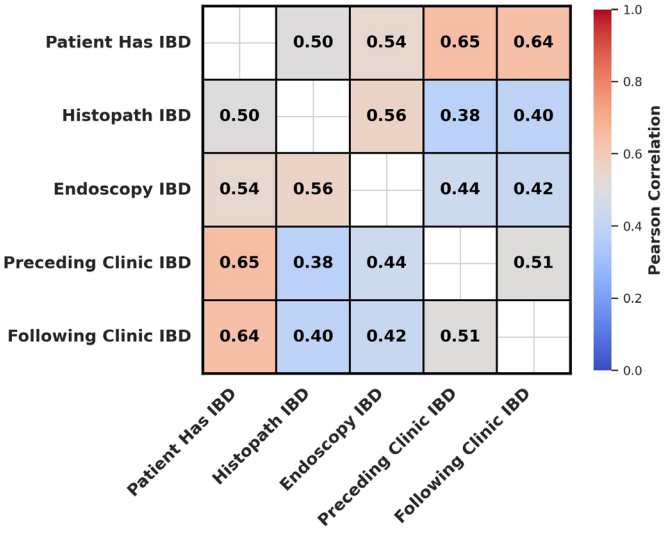

Figure 2 Pearson correlations between document-level and patient-level inflammatory bowel disease (IBD) diagnosis. In some cases correlations are surprisingly low (0.38-0.44).

The training set contained 1128 (70%) patients, and the test set contained 484 (30%). Within the training set, 809 (72%) patients had IBD, and 319 (28%) did not. Within the test set, 351 (73%) of the patients had IBD, and 133 (27%) did not.

In total, 9311 free-text documents were manually reviewed. The training set contained 6559 documents, of which 4290 (65%) were labelled as suggestive of IBD and 2269 (35%) were not. The test set contained 2752 free-text documents, of which 1725 (63%) were labelled as suggestive of IBD and 1027 (37%) were not, after removing all microscopic colitis cases. There were 2592 rows of carefully aligned document data in the final data set—1824 in the training set and 768 in the test set. Coverage is reported according to the number of total rows used. There were no significant differences between the development and test settings, eligibility criteria, outcome and predictors.

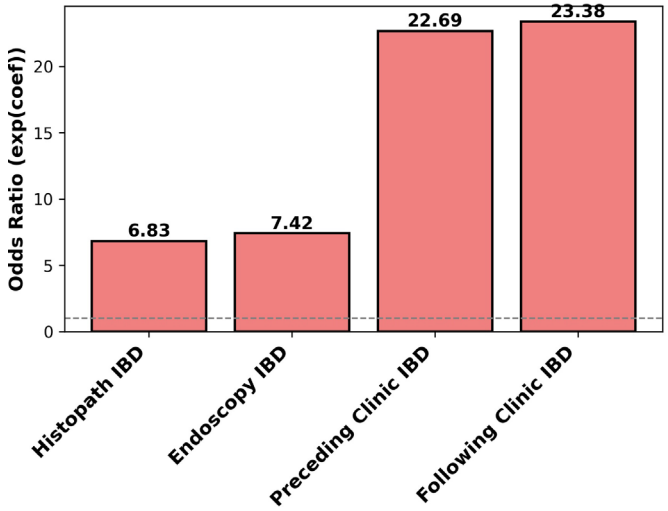

Figure 3 Odds Ratios (OR) by document type. Highlights the different weights of positive document identification as a contributor to patient-level inflammatory bowel disease (IBD) prediction.

Cohort demographics

No significant differences were observed between the training and test cohorts in any continuous demographic. However, the test cohort had ~3% more females, and it was ~2% less white—these were the only two significant results. Full demographic results are provided in online supplemental appendix F.

UMLS versus free text

The addition of UMLS had a mildly detrimental effect on overall performance by an average of 1% precision. 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 can all map to IBD-associated concept unique identifiers within UMLS. The UMLS Preferred Terms (PTs) are thus vulnerable to overmapping across ontologies, making its usage non-justifiable for this task, especially given the additional overhead added for users. Accordingly, the rest of the study focuses only on the use of raw free-text NLP models.

Document-level full results

At a document level, model performance was variable. Top performers included DistilBERT (micro F1: 93.92%) and sBERT (micro F1: 93.75%). Full results are given in online supplemental appendix G. Larger LLMs performed best overall in terms of specificity (79.43–83.64%) with comparable MCC scores (0.6602–0.7131) to the best BERT-based models, suggesting that in terms of performance, these models overall have the performance edge in document classification because they had not seen the training data before.

Patient-level full results

At a patient level, model performance degrades. In particular, average recall (−0.97%), NPV (−2.93%), and Brier (+0.031) scores suffer. When grouped by type, the fully trained and fine-tuned model's performance all degrades slightly overall as per table 1. In contrast, LLMs experience a sizeable average deterioration, particularly in NPV (−11.9%) and recall (−2.2%) when moving from document to patient level, even if precision (+3.47%) and specificity (+1.09%) improve slightly. Reseeding the test set did not drastically alter the results.

In particular, the 32B LLMs' Brier scores substantially worsened when moving to patient-level prediction, unlike the 70B model, which remained comparatively stable. Almost all the models underpredicted at lower probabilities and overpredicted at higher probabilities, as highlighted in online supplemental appendix H. As a general rule, any model with a Brier score above 0.12 suffers from substantial calibration issues, with sBERT-Base being the overall best calibrated according to the Brier scores (*Document-Level: 0.0706, Patient-Level: 0.1095*) and plotting (online supplemental appendix H).

Table 2 Full economic results by model

Model	Training + Inference Time (minutes)	Total kWh	CO2 Emissions (grams)
Regex	1.56	0.005	1.04
spaCy	12.48	0.030	6.21
BoW	1.46	0.00786	1.63
TF-IDF	1.48	0.0079	1.64
Word2Vec	33.65	0.18	37.36
sBERT	5.61	0.045	9.31
sBERT Med	2.68	0.019	3.96
DistilBERT	119.71	1.11	230.09
BioclinicalBERT	213.55	2.01	416.73
RoBERTa	228.29	2.13	441.75
Mistral-0.3-7B*	1075	4.54	938.50
M42-Llama_8B*	1080	4.55	942.90
DeepSeek-R1- Qwen2.5_32B*	4142	27.75	5745.90
Qwen3_32B*	4146	27.78	5751.50
DeepSeek-R1-Llama70B*	18 050	163.65	33 884.40

Full economic and sustainability analysis results for the included models.
*No training—inference only.
BOW, bag of words; regex, regular expressions; SBERT, sentence-bidirectional encoder representations from transformers; Spacy, negation strategies; TF-IDF, term frequency-inverse document frequency; Word2Vec, word-2-vector.

Document-level and patient-level interactions

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 (0.64–0.65) with patients ultimately having IBD than endoscopy reports (0.54) or histopathology reports (0.50).

Document to patient regression and tree models

L2 (ridge) based LR models were used to assess the predictive performance of document types towards patient-level IBD diagnosis. These full results are given in online supplemental appendix H, including the results of the single versus multidocument per patient comparison. Clinic letters hold more predictive weight (OR: 22.69–23.38) than the other factors as per figure 3.

Finally, a DT 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 clinical splits represented the first branch, suggesting that document-to-patient level mapping is not simply solved. The visual tree is shown in online supplemental appendix H, along with details of document type performance variation.

Fairness analysis

In online supplemental appendix I, the full fairness results are given for each model at a patient level. At baseline analysis, bias was identified as a significant problem for all the more effective model types. Overall, the best performing locally trained models (BoW, TF-IDF, sBERT, DistilBERT and BioClinicalBERT) were probably biased against females (DP: 0.104 to 0.134) the wealthy (DP: ¯0.108 to ¯0.174, DI: 0.791 to 0.866) and those of African ethnicity (DP: 0.128 to 0.263, EO: 0.031 to 0.176, DI: 1.229 to 1.266). LLMs, in contrast, were biased against older patient groups (EO: 0.105 to 0.215), the wealthy (DP: ¯0.123 to 0.234, EO: 0.023 to 0.169, DI: 0.670 to 1.014) and slightly towards those of African ethnicity (EO: 0.52 to 0.118) with the sole exception of m42 which was biased against African patients (DP: 0.125). RB models had no particular biases, but were also ineffective. Examining underlying fundamental differences in the cohort reinforced the assertion that, in particular, the model gender biases are real because female patients had an OR of 0.91 for IBD overall at baseline. Wealth-related biases are, however, likely related to true differences in the underlying cohorts. Ethnicity biases are more problematic to be confident about because there were only 9 African patients in the test cohort. LLMs appear to have higher recall for older patients having IBD (EO: 0.105 - 0.215) than all other model types, which don't appear to discriminate in this fashion (EO: 0.000 - 0.048). This suggests a greater degree of intelligence among the LLMs, as the 60-70 age group in the cohort had an OR of 1.15 to have IBD compared to the 20-30 year olds - something unexpected that the BERT models did not pick up on.

Economic comparison

A complete economic analysis was undertaken by model, as per table 2.

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 far better results overall. Regex, BoW and TF-IDF were the quickest/cheapest overall, but once calibration is considered as well, sBERT outperforms them. Running inference with LLMs is slow, and, in the case of the largest models, it is currently prohibitively costly.

Explainability analysis

The string-based methods are entirely explainable. SHAP plots for BoW, TF-IDF, Word2Vec and sBERT all highlight similar word token patterns, as highlighted in online supplemental appendix J. Additionally, in this appendix some examples of model-breaking (and fixing) text fragments are given, which help to highlight how these models are making decisions and illustrate visually where some of the flaws lie and the degree to which these models are currently overfitted and need further finetuning by others to become truly generalisable. LLM explainability is more complex, requiring a bespoke approach that will be reported on in a subsequent study.

DISCUSSION

This study has provided clinicians with a set of open weight models (https://huggingface.co/collections/MattStammers/a-collection-of-ibd-bert-models-682b01badbaa646380f54b14) of high quality, which can be used out of the box or, ideally, further finetuned to start reidentifying patients with IBD from clinical free-text. However, the study also emphasises that patient-level cohort identification is somewhat context-dependent and cannot be conducted in isolation. This unexpected complication was mentioned by Schmidt *et al*[41] last year. However, this study goes further, serving as an in-depth example of the added complexities of chronic disease cohort extraction,[3] not encountered with, for instance, adenoma detection.[5 42] This study also highlights that endoscopic and histopathological IBD diagnoses made in isolation are quite often wrong, and accurate IBD diagnoses can only be made when the whole context is made available.

LLMs performed surprisingly well at this task, especially considering that they were never pre-trained on the data sets involved. In particular, they retained high precision and specificity for the task, even if recall and NPV often lagged, and costs were high. sBERT, DistilBERT, and BioclinicalBERT exhibited strong performance characteristics, but are likely somewhat overfitted to the training data. Simple RB methods have extremely high recall and are very fast, but suffer from very low specificity. Hybrid pipelines could be established using rule, ML or BERT-based classifiers for pre-screening, followed by LLMs for final patient selection; however, this approach would take substantial further work to perfect in light of the complex document-patient interactions identified already. It is likely that the costs of LLMs will decrease rapidly while their usefulness will only continue to increase.

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] In addition, the validation of the gold standard cohort in this study was robust and led by a senior gastroenterologist with strong informatics experience. Rerunning the experiment with and without microscopic colitis and testing UMLS alongside a variety of evaluation metrics provides additional confidence in the results. Future work can build on these open-source resources (https://github.com/MattStammers/An_Open_Source_Collection_Of_IBD_Cohort_Identification_Models) to improve future generalisability and quality.

The weaknesses of this study include bias in the locally trained models, which reflect the biases inherent in the local training data. Other weaknesses include class imbalance within the cohorts, although removing the patients with microscopic colitis improved this, as did including the F1 score and MCC as primary evaluation metrics. Additionally, the way the IBD cohorts were selected resulted in model calibration problems as seen in the prior study,[3] although in some cases (eg, sBERT-Base), these effects were significantly ameliorated. Another weakness of the study is its single-site nature. Conducting multisite studies like this is not easy at present due to the high costs of cloud computing, which will hopefully soon reduce.[43] 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 models would have performed significantly better, the evidence we have suggests that large LLMs are sometimes even less faithful when managing factual information, even if readability and informativeness improve.[44] Although the 70B model did outperform the 32B models on some parameters, both the 32B models exhibited higher specificity. Because of its comparatively high carbon footprint, this model was only run over the validation set and just over 50% of the total data set. While it is possible that a multi-classification model including 'possible IBD' as a category might help improve performance on the local test cohort, it is doubtful that this would improve performance in reality, as it would result in arbitrary class splitting and would also reduce clinical clarity, while likely increasing model brittleness.

A simple demo app (https://huggingface.co/spaces/MattStammers/IBD_Cohort_Identification) is made available for users to demonstrate how easy the models are to use. The models should function with any English-language clinical free text without substantial preprocessing required, and they should operate seamlessly with either the transformers or scikit-learn libraries as appropriate. The LLM prompt templating system (online supplemental appendix C), built on CARP, has been shown to clinically generalise in this study and would likely work for the classification of other chronic diseases. Further work on prompt engineering and its effects on

improving model performance in this field, along with exploring LLM explainability in depth, is planned.

CONCLUSION

NLP has undergone significant advancements over the last 50 years. LLMs appear likely to solve the problem of re-identifying patients with IBD from clinical free text sources in the future. Once cost, performance and bias issues are addressed, they and their successors are likely to become the primary method of AI data retrieval and clinical data warehousing.

All models and weights from this study are released as open source/weight (https://github.com/MattStammers/An_Open_Source_Collection_Of_IBD_Cohort_Identification_Models) to suit all environments and preferences, including resource-constrained environments.

Author affiliations
[1]Department of Gastroenterology and The Southampton Emerging Therapies and Technologies (SETT) Centre, University Hospital Southampton NHS Foundation Trust, Southampton, UK
[2]University of Southampton Faculty of Medicine, Clinical Informatics Research Unit (CIRU), University of Southampton, Southampton, UK
[3]University Hospital Southampton NHS Foundation Trust, Southampton, UK
[4]University of Southampton, Southampton, UK
[5]Queen's Medical Centre, Nottingham, UK
[6]University of Southampton School of Healthcare Enterprise and Innovation, Southampton, UK

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

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. AI was not used to write the document. It was only used in the experiments and debugging of the code, which is provided open source. The author accepts full responsibility for the code, analytics and results.

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 Programme. Study sponsorship was provided by the UHS Research and Development (R&D) Department. The protocol was independently developed.

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. The remaining authors declare no competing interests.

Patient consent for publication Not applicable.

Ethics approval The Wessex REC and HRA provided research ethics board approval for this study (IC-IBD -23/SC/0152) on 16 May 2023 (https://www.hra.nhs.uk/planning-and-improving-research/application-summaries/research-summaries/ic-ibd-ibd-cohort-identification-study/).The ethics committee waived the requirement for informed consent on the basis that the information was for the patient's benefit and was carefully managed following best information governance and cybersecurity practices.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information. Source data sharing is not possible in this study because the data were not collected for this purpose; however, additional secondary data can be made available upon request. 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_Identification_Models. Models are made fully accessible, open source at: https://huggingface.co/collections/MattStammers/a-collection-of-ibd-bert-models-682b01badbaa646380f54b14 and in the GitHub repo (https://github.com/MattStammers/An_Open_Source_Collection_Of_IBD_Cohort_Identification_Models). The LLMs are all open source and can be accessed at the following links: (1) Mistral 7b-v0.3-Instruct[45]: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3; (2) Llama3-Med42-8b[46]: https://huggingface.co/m42-health/Llama3-Med42-8B; (3) Deepseek-Qwen2.5-32B[47]: https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B; (4) Qwen-v3-32[48]: https://huggingface.co/Qwen/Qwen3-32B; (5) Deepseek-Llama-70B[49]: https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B.

ORCID iD
Matt Stammers http://orcid.org/0000-0003-3850-3116

REFERENCES

1 Baumgart DC, Sandborn WJ. Inflammatory bowel disease: clinical aspects and established and evolving therapies. *Lancet* 2007;369:1641–57.
2 Nikolaus S, Schreiber S. Diagnostics of inflammatory bowel disease. *Gastroenterology* 2007;133:1670–89.
3 Stammers M, Sartain S, Cummings JRF, *et al*. Identification of Cohorts with Inflammatory Bowel Disease Amidst Fragmented Clinical Databases via Machine Learning. *Dig Dis Sci* 2025;13:1–4.
4 Wornow M, Lozano A, Dash D, *et al*. Zero-Shot Clinical Trial Patient Matching with LLMs. *NEJM AI* 2025;2:AIcs2400360.
5 Stammers M, Ramgopal B, Owusu Nimako A, *et al*. A foundation systematic review of natural language processing applied to gastroenterology & hepatology. *BMC Gastroenterol* 2025;25:58.
6 Salton G, Wong A, Yang CS. A vector space model for automatic indexing. *Commun ACM* 1975;18:613–20.
7 McCallum A, Nigam K. A comparison of event models for naive bayes text classification. 1998. Available: https://www.semanticscholar.org/paper/A-comparison-of-event-models-for-naive-bayes-text-McCallum-Nigam/04ce064505b1635583fa0d9cc07cac7e9ea993cc
8 Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Inf Process Manag* 1988;24:513–23.
9 Turney PD, Pantel P. From Frequency to Meaning: Vector Space Models of Semantics. *J Artif Intell Res* 2010;37:141–88.
10 Deerwester S, Dumais ST, Furnas GW, *et al*. Indexing by latent semantic analysis. *J Am Soc Inf Sci* 1990;41:391–407.
11 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Adv Neural Inf Proc Syst, Curran Associates, Inc; 2017. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
12 Devlin J, Chang MW, Lee K, *et al*. BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv* 2019. Available: http://arxiv.org/abs/1810.04805
13 Radford A, Narasimhan K, Salimans T, *et al*. Improving language understanding by generative pre-training. *Comp Sci Linguistics* 2018.
14 Sanh V, Debut L, Chaumond J, *et al*. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv* 2020. Available: http://arxiv.org/abs/1910.01108 https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035
15 Liu Y, Ott M, Goyal N, *et al*. RoBERTa: a robustly optimized BERT pretraining approach. *arXiv* 2019. Available: http://arxiv.org/abs/1907.11692
16 Alsentzer E, Murphy JR, Boag W, *et al*. Publicly available clinical BERT embeddings. *arXiv* 2019. Available: http://arxiv.org/abs/1904.03323

17 Floridi L, Chiriatti M. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds Mach* 2020;30:681–94.
18 Nori H, Lee YT, Zhang S, *et al*. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv* 2023. Available: http://arxiv.org/abs/2311.16452
19 Jiang AQ, Sablayrolles A, Roux A, *et al*. Mixtral of experts. *arXiv* 2024. Available: http://arxiv.org/abs/2401.04088
20 Christophe C, Kanithi PK, Raha T, *et al*. Med42-v2: a suite of clinical LLMs. *arXiv* 2024. Available: http://arxiv.org/abs/2408.06142
21 DeepSeek-AI GD, Yang D, Zhang H, *et al*. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv* 2025. Available: http://arxiv.org/abs/2501.12948
22 Zheng X, Li Y, Chu H, *et al*. An empirical study of Qwen3 quantization. *arXiv* 2025. Available: http://arxiv.org/abs/2505.02214
23 Collins GS, Moons KGM, Dhiman P, *et al*. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024;385:e078378.
24 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;162:55–63.
25 Sun X, Li X, Li J, *et al*. Text classification via large language models. *arXiv* 2023. Available: http://arxiv.org/abs/2305.08377
26 Pate A, Riley RD, Collins GS, *et al*. Minimum sample size for developing a multivariable prediction model using multinomial logistic regression. *Stat Methods Med Res* 2023;32:555–71.
27 Juckett D. A method for determining the number of documents needed for a gold standard corpus. *J Biomed Inform* 2012;45:460–70.
28 Rufibach K. Use of Brier score to assess binary predictions. *J Clin Epidemiol* 2010;63:938–9.
29 Goel N, Yaghini M, Faltings B. Non-discriminatory machine learning through convex fairness criteria. AAAI Conf Art Int. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11662
30 Hardt M, Price E, Price E, *et al*. Equality of opportunity in supervised learning. Adv Neural Inf Proc Syst, Curran Associates, Inc; 2016. Available: https://proceedings.neurips.cc/paper_files/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html
31 Zafar MB, Valera I, Gomez Rodriguez M, *et al*. Fairness beyond disparate treatment & disparate impact: learning classification without disparate mistreatment. Proceedings of the 26th international conference on world wide web. Republic and Canton of Geneva, CHE: International world wide web conferences steering committee, (WWW '17); 2017:1171–80. Available: https://doi.org/10.1145/3038912.3052660
32 CodeCarbon — CodeCarbon 3.0.1 documentation. Available: https://mlco2.github.io/codecarbon/ [Accessed 12 May 2025].
33 GOV.UK. Greenhouse gas reporting: conversion factors 2024. 2024. Available: https://www.gov.uk/government/publications/greenhouse-gas-reporting-conversion-factors-2024 [Accessed 13 May 2025].
34 Mosca E, Szigeti F, Tragianni S, *et al*. SHAP-based explanation methods: a review for NLP interpretability. In: Calzolari N, Huang CR, Kim H, et al, eds. *Proc 29th Int Conf Comp Ling. Gyeongju, Republic of Korea: international committee on computational linguistics*. 2022: 4593–603. Available: https://aclanthology.org/2022.coling-1.406/
35 Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. *Knowl Inf Syst* 2014;41:647–65.
36 Ribeiro MT, Singh S, Guestrin C. 'Why should I trust you?': explaining the predictions of any classifier. *arXiv* 2016. Available: http://arxiv.org/abs/1602.04938
37 Salih AM, Raisi-Estabragh Z, Galazzo IB, *et al*. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Adv Intell Syst* 2025;7:2400304.
38 Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-Networks. *arXiv* 2019. Available: http://arxiv.org/abs/1908.10084
39 Deka P, Jurek-Loughrey A, Deepak P. Evidence extraction to validate medical claims in fake news detection. In: Traina A, Wang H, Zhang Y, *et al*., eds. *Health Inf Sci*. Cham: Springer Nature Switzerland, 2022: 3–15.
40 Gururangan S, Marasović A, Swayamdipta S, *et al*. Don't stop pretraining: adapt language models to domains and tasks. *arXiv* 2020. Available: http://arxiv.org/abs/2004.10964
41 Schmidt L, Ibing S, Borchert F, *et al*. Automating clinical phenotyping using natural language processing: an application for crohn's disease. *medRxiv* 2024.
42 Shivade C, Raghavan P, Fosler-Lussier E, *et al*. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* 2014;21:221–30.
43 NHS Transformation Directorate. Secure data environments (SDEs). Available: https://transform.england.nhs.uk/key-tools-and-info/data-saves-lives/secure-data-environments/ [Accessed 22 Sep 2023].
44 Mahapatra J, Garain U. Impact of model size on fine-tuned LLM performance in data-to-text generation: a state-of-the-art investigation. *arXiv* 2024. Available: http://arxiv.org/abs/2407.14088
45 Hugging Face. mistralai/Mistral-7B-Instruct-v0.3, Available: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3 [Accessed 21 May 2025].
46 Hugging Face. m42-health/Llama3-Med42-8B, 2024. Available: https://huggingface.co/m42-health/Llama3-Med42-8B [Accessed 21 May 2025].
47 Hugging Facecc. deepseek-ai/DeepSeek-R1-Distill-Qwen-32B. 2025. Available: https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B [Accessed 21 May 2025].
48 Hugging Face. Qwen/Qwen3-32B. 2025. Available: https://huggingface.co/Qwen/Qwen3-32B [Accessed 21 May 2025].
49 Hugging Face. deepseek-ai/DeepSeek-R1-Distill-Llama-70B, 2025. Available: https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B [Accessed 21 May 2025].