Intelligent Multi-Label Triage for Hematological Comorbidities using BERT

Jan Banasik[1][0009-0007-0775-9924] and Antoni Pater[1][0009-0000-8347-097X]

[1] AGH University of Krakow
Department of Applied Computer Science
al. A. Mickiewicza 30, 30-059 Krakow, Poland
{jbanasik, antonipater}@student.agh.edu.pl

Abstract. Modern triage requires simultaneous evaluation of multiple organ systems; single-label classifiers fail to capture comorbidity. We address clinical routing from basic lab parameters and a minimal adaptive interview triggered when a value is out of reference range. We posit that a BERT-based model pre-trained in the medical domain (masked language modelling, MLM) can route patients to eight classes (primary care, specialists, emergency care). We use a hybrid data approach: MIMIC-III/IV (specialists and emergency cases) and Synthea (chronic trajectories and primary care). Labels come from ICD-to-specialist mapping; lab results are quantised for triage. We focus on methodology: MLM pre-training on the hybrid corpus, multi-label fine-tuning with a clinically cost-sensitive loss, and honest patient-level validation. Results include ROC curves, calibration (expected calibration error, ECE), precision–recall curves, confusion matrices, and attention visualisation for interpretability. High ROC AUC for emergency routing and low ECE support human-in-the-loop deployment with emphasis on safety.

Abbreviations. Standard ML/clinical NLP terms include BCE, ECE, EHR, ICU, LSTM, MLM, NER, NLP, RNN, ROC, SVM, XAI, AUC, AUPRC, TPR, FPR, FN, and FP. POZ: *podstawowa opieka zdrowotna* (Polish primary care). SOR: *szpitalny oddział ratunkowy* (Polish ED).

1 Introduction

Demographic ageing has led to a sharp rise in patients with multiple chronic conditions. In acute and primary care, triage and referral decisions must often be made on the basis of limited history and a small set of tests—for example basic blood count and biochemistry—before a full workup is available. At the same time, missing an urgent case (e.g. severe anaemia or acute kidney injury) or failing to flag the need for several specialists at once can delay treatment and worsen outcomes. We therefore prioritise safety: any decision-support system should assist, not replace, the physician, and must correctly identify cases requiring emergency care (SOR; Polish ED) or multi-specialist input.

Traditional triage tools rely on rule-based logic or single-label classifiers. Rule-based systems become hard to maintain when symptom descriptions are ambiguous or when the volume of free-text notes grows; they also tend to use fixed thresholds on lab values, which ignore patient-specific baselines and trends. Single-label models assign each case to exactly one class and are structurally unsuited to comorbidity: increasing the probability of a cardiology referral forces a decrease in the probability of a pulmonology referral, even when both are clinically indicated. The triage function should output a multi-label vector—independent probabilities for several specialist queues—so that multiple referrals can be activated when the clinical picture warrants it.

In this article we posit that a BERT-based model pre-trained in the medical domain using MLM can appropriately route patients to multiple classes at once, with interpretability provided by attention visualisation. Our objective is to propose and validate a pipeline that combines (1) quantisation of laboratory parameters into semantic tokens, (2) a minimal adaptive interview triggered when a value falls outside the reference range, and (3) multi-label BERT fine-tuned with a loss that reflects clinical costs (e.g. higher cost for missing SOR than for a false alarm). We define each interview item in a reproducible way (parameter, reference range, question from a bank, answer token), so that the flow from lab result to routing suggestion is deterministic and auditable. We focus on the methodology and results of this model, without comparison to other classifiers.

Clinical text classification and triage have a long history. Early systems relied on rule-based keyword extraction and hand-crafted thresholds; later, SVMs and logistic regression were applied to bag-of-words or simple feature representations. These approaches often ignored negation, temporal order, and the interplay between structured data and free text. With the rise of deep learning, RNNs and LSTMs became the standard for modelling clinical sequences [1]; the transformer [2] replaced recurrence with self-attention, and BERT [3] pre-trains encoders with MLM, then fine-tunes them for downstream tasks. In the biomedical domain, variants such as BioBERT and ClinicalBERT have been used mainly for NER and single-label phenotyping.

Relation to prior work. Most clinical NLP work using BERT focuses on single-label tasks or on NER rather than triage. Our pipeline differs in three ways: (i) multi-label output with sigmoid (no single-class constraint), (ii) in-domain MLM pre-training on the same hybrid corpus used for fine-tuning, and (iii) an adaptive interview driven by reference ranges and a question bank. We do not compare to other classifiers in this paper; the focus is on methodology and validation of this design, in line with similar single-methodology studies; future work may include literature baselines and ablations.

To our knowledge, few works address multi-label comorbidity triage with a BERT-style model trained on a hybrid corpus (Synthea + MIMIC) coupled with an adaptive interview defined by reference ranges and a question bank. Our contribution is to combine in-domain MLM pre-training, multi-label fine-tuning with a clinically cost-sensitive loss, and strict patient-level validation, and to

report discriminatory performance (ROC, precision–recall), calibration (ECE), confusion analysis, and attention-based interpretability.

The remainder of the paper is organised as follows. Section 2 describes the hybrid corpus and the adaptive interview. Section 3 presents the model architecture, pre-training and fine-tuning, and the loss function. Section 4 reports experiments and results. Section 5 summarises findings, limitations, and future work.

2 Data and Methodology

The corpus is a hybrid dataset we constructed for training and evaluating the triage model. It merges two sources. Synthea [5] is an open-source synthetic patient generator that simulates longitudinal EHRs with chronic conditions (e.g. anaemia, diabetes, chronic kidney disease); its patterns may be more regular than in real clinical practice. MIMIC-III/IV [4] is a de-identified critical care database on PhysioNet; it contains real ICU data, including acute presentations and narrative clinical notes. Combining both allows the model to see both chronic, multi-morbid trajectories and acute, life-threatening events, reducing overfitting risk.

Records were preprocessed into a single token format. Structured laboratory values were quantised into semantic tokens based on reference ranges (e.g. `HGB_LOW`, `CREATININE_HIGH`, `PLT_CRITICAL_LOW`), so that the model can learn non-linear risk thresholds rather than raw numbers. Free-text symptoms and diagnoses were mapped to symptom or diagnosis tokens. Each sample is a sequence of such tokens plus a multi-label binary vector of length eight: primary care (POZ), Gastroenterology, Haematology, Nephrology, Emergency (SOR), Cardiology, Pulmonology, Hepatology. Labels are derived from ICD-to-specialist mapping; a single patient can have multiple labels. The corpus was split with strict patient-level separation: every encounter of a given patient belongs to one set only, so the model never sees the same patient in both training and validation. Class distribution is imbalanced (e.g. SOR rarer); the cost-sensitive loss and per-class thresholds handle this. Table 1 summarises the dataset.

Tokens such as `HGB_Q5` or `CREATININE_Q9` denote ten discrete bands $Q1$–$Q10$ defined heuristically from reference and critical bounds (below norm, within norm split into sub-bands, above norm, and critical tails), not population quantiles. For example, $Q4$–$Q7$ subdivide the in-range interval; $Q9$ marks a strongly elevated value below the critical band. This naming is inherited from the preprocessing code and should not be read as "fifth quintile" or "ninth decile" in the statistical sense.

Example. A training sample is a space-separated token sequence, e.g. `AGE_60 SEX_M HGB_Q5 HCT_Q5 PLT_Q6 MCV_Q5 WBC_Q5 CREATININE_Q9 ALT_Q5` (male, 60; HGB/HCT in mid in-range bands $Q5$; platelet count $Q6$; creatinine in the high above-range band $Q9$). The multi-label vector has value 1 for Nephrology when that encounter is labelled for nephrology referral only. Sequences are truncated or padded to 128 tokens; the model sees `[CLS]` $\oplus$ token sequence $\oplus$ `[SEP]`.

Table 1. Dataset characteristics (hybrid corpus: Synthea + MIMIC).

Item	Value
Training samples (encounters)	260 663
Validation samples (encounters)	65 166
Vocabulary size (tokens)	143
Classes	8
Patient-level split	80% / 20%

2.1 Adaptive Interview

In deployment, when a value falls outside the reference range (e.g. low haemoglobin), the system can ask a follow-up question (e.g. "Any recent bleeding?") before routing. We call this an adaptive interview: it is triggered only when a laboratory value is out of range, and the question is chosen from a bank according to the parameter and the direction of the deviation. Reference ranges are defined in `lab_norms.json` (can depend on age and sex). Quantisation maps each lab value to a discrete token (e.g. `HGB_Q1`, `CREATININE_Q9`, `CREATININE_Q10_CRITICAL`), allowing the model to learn non-linear risk boundaries without regressing on raw numbers. For reproducibility and audit, we define an interview item as the minimal logged unit: (1) triggering parameter and quantised token, (2) applied reference range, (3) question from `questions.json`, (4) answer encoded as `token_yes` or `token_no` appended to the BERT input. The path from lab result → question → answer → routing suggestion is thus deterministic and auditable.

2.2 Model Architecture

The input is a sequence of tokens (demographics, quantised lab values, composite features, symptoms, optional interview answers). These are embedded and passed through a BERT encoder pre-trained with MLM, then fine-tuned for multi-label classification. The `[CLS]` representation is fed into a linear layer with eight outputs, each passed through sigmoid (not softmax), so the model outputs eight independent probabilities in $[0, 1]$. Figure 1 illustrates this. The implementation and training are done in PyTorch.

Tokenisation. A word-level tokeniser is built on the training corpus; vocabulary size ∼140. Special tokens: a dedicated padding id (index 0 in the saved vocabulary), `[UNK]`, `[CLS]`, `[SEP]`, `[MASK]`. Input format: `[CLS]` ⊕ token sequence ⊕ `[SEP]`, length 128 (age/sex, quantised labs, composite risk, symptoms, interview-answer tokens).

Pre-training (MLM). The BERT encoder is trained from scratch on the same corpus with the standard MLM objective [3]: 15% of positions masked (or replaced/unchanged with small probability), model predicts original tokens. This adapts representations to the medical domain and token vocabulary. Configuration: 6 layers, hidden 256, 8 heads, FFN 1024; AdamW lr 10^{-4}, batch 64, up to 15 epochs.

Fine-tuning. The `[CLS]` representation is passed through a linear layer (8 outputs) and sigmoid, so the model outputs eight independent probabilities

(multi-label). We use BCEWithLogitsLoss; training: batch 32 (train), 64 (val), lr 2×10^{-5}, up to 5 epochs. Per-class decision thresholds are tuned on the validation set (e.g. Youden's index or F1 per class), with a lower threshold for SOR to prioritise sensitivity. Table 2 summarises hyperparameters.

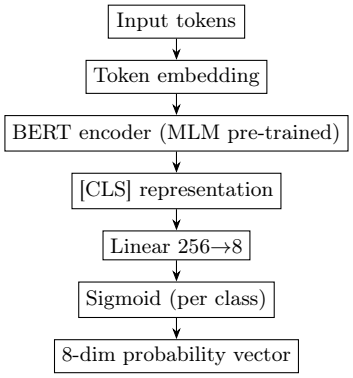

Fig. 1. Model architecture: input → embedding → BERT (MLM pre-trained) → [CLS] → linear → sigmoid → eight class probabilities.

Table 2. Model and training configuration.

Component	**Setting**
Encoder (pre-training)	6 layers, hidden 256, 8 heads, FFN 1024
Vocabulary size	∼140 tokens
Max sequence length	128
MLM pre-training	AdamW, lr 10^{-4}, batch 64, up to 15 epochs
Fine-tuning	AdamW, lr 2×10^{-5}, batch 32 (train), 64 (val), 5 epochs
Classification head	Linear → Sigmoid, 8 classes
Loss	BCEWithLogitsLoss + cost-sensitive focal loss
Train/val split	80% / 20%, patient-level disjoint
Framework	PyTorch

2.3 Loss Function and Clinical Costs

The base loss is BCEWithLogitsLoss. We extend it to a cost-sensitive focal loss. A clinical cost matrix C_{ij} defines the cost of classifying true class i as class j: missing SOR (false negative) has high cost (e.g. 8–10); false positive for SOR is less costly. Confusing Haematology with Nephrology may have moderate cost (e.g. 7). The loss is weighted so that higher-cost errors contribute more to the gradient. Focal loss ($\gamma > 0$) downweights easy examples and upweights hard ones (borderline and rare cases). This prioritises safety-critical errors (e.g. missing SOR) and difficult cases during training.

2.4 Honest Validation and Reproducibility

We split data at the patient level: all encounters of the same `subject_id` go to the same set, so the model never sees the same patient in both training and

validation (no patient-level leakage). For Synthea we use the synthetic patient identifier. We approximately respect the multi-label distribution so validation metrics are representative. Synthea is rule-based and may be more regular than real data; the hybrid corpus and validation set containing both synthetic and real patients reduce overfitting risk to the generator. Preprocessing and training are in Python; configuration in YAML/JSON. MIMIC data were used under PhysioNet credentialing. The system is decision support only; the physician remains responsible for the final decision. Deployment would require local ethics and site-specific validation.

3 Experiments and Results

We evaluate on the held-out validation set (20% after patient-level split). We report accuracy, precision, recall, F1 (macro/weighted), ROC AUC per class (one-vs-rest), AUPRC, and ECE. Table 3 summarises key metrics per class.

3.1 Summary and Deployment Implications

High ROC AUC for SOR and for Haematology and Nephrology means the model separates positive from negative cases well. Low ECE indicates that predicted probabilities can be shown to the physician. They do not exhibit systematic over- or under-confidence. Confusion matrices (SOR, Haematology, Nephrology) show high sensitivity for SOR and diagonal dominance for the specialist classes. Cases near the decision threshold or with conflicting specialist flags can be routed to dual review or trigger additional interview questions; attention visualisation helps explain which tokens drove the suggestion.

3.2 ROC and Precision–Recall

Figure 2 plots ROC curves (TPR vs FPR) for each of the eight classes (top) and Precision-Recall curves (bottom). The SOR curve reaches the top-left corner (ROC AUC ≈ 1.0), indicating effective separation of urgent cases. Nephrology and Haematology also show high AUC. Lower AUC for Pulmonology reflects overlap with Cardiology (e.g. dyspnoea and chest pain). SOR and Haematology have high AUPRC, supporting deployment with clinician oversight.

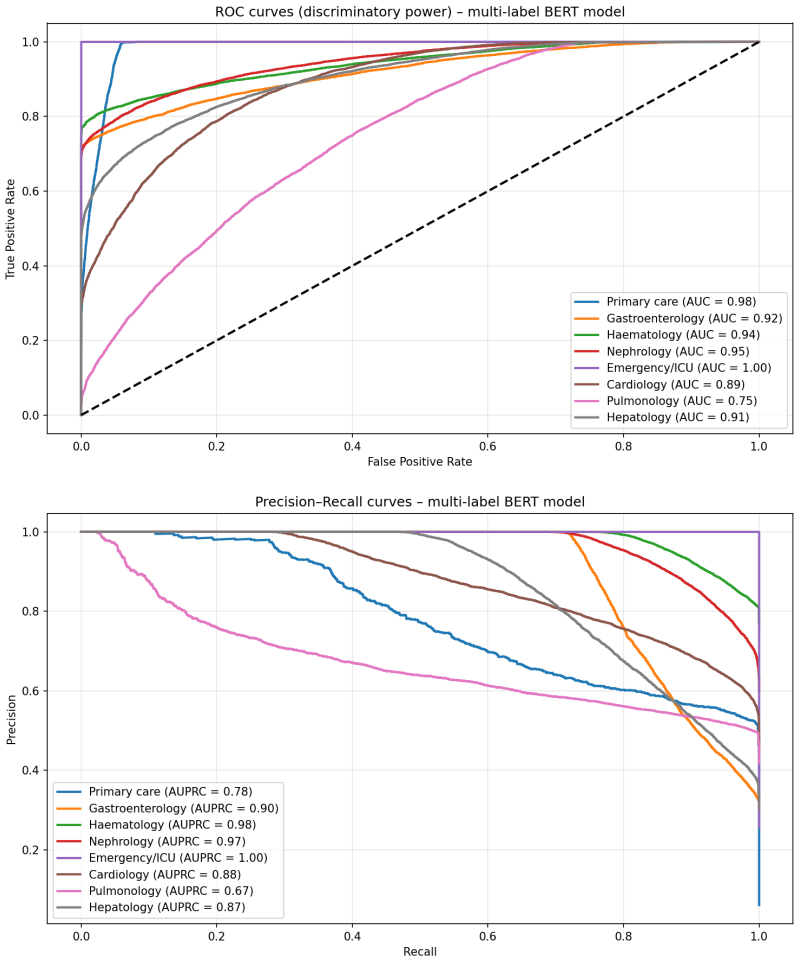

Fig. 2. Discriminatory power: (top) ROC curves; (bottom) precision–recall curves (multi-label BERT model).

Table 3. Example classification metrics per class. ROC AUC = area under the ROC curve; ECE = expected calibration error.

Class	ROC AUC	F1	ECE
POZ (primary care)	0.98	0.683	0.041
Gastroenterology	0.92	0.836	0.003
Haematology	0.94	0.913	0.011
Nephrology	0.95	0.883	0.012
SOR (emergency)	1.00	1.000	0.000
Cardiology	0.89	0.782	0.026
Pulmonology	0.75	0.669	0.066
Hepatology	0.91	0.753	0.072

3.3 Confusion Matrices and Calibration

Figure 3 shows 2×2 confusion matrices for SOR, Haematology, and Nephrology (one-vs-rest). SOR has high sensitivity (urgent cases not missed); Haematology and Nephrology show strong diagonal counts. Figure 4 shows calibration curves: horizontal axis = mean predicted probability per bin, vertical = empirical accuracy. Our model achieves ECE in the range 0.007–0.012 across classes, well below 0.05; curves lie close to the diagonal, so confidence estimates can be trusted in deployment.

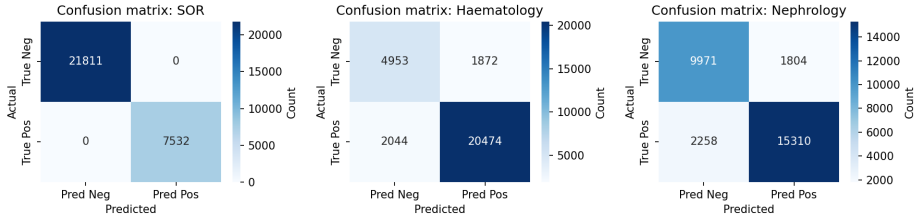

Fig. 3. Confusion matrices (one-vs-rest): (a) SOR (Emergency), (b) Haematology, (c) Nephrology.

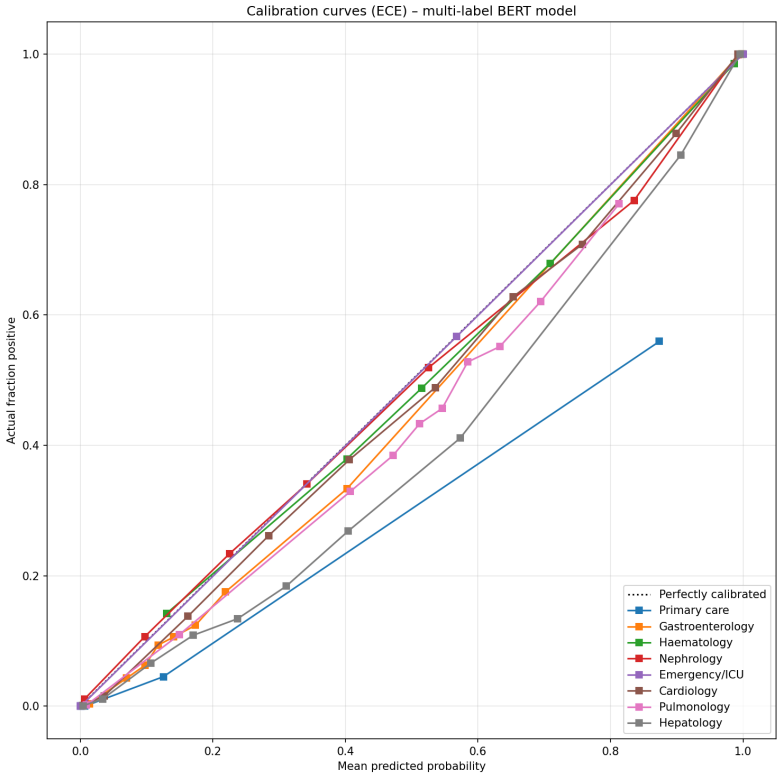

Fig. 4. Calibration curves (ECE)—multi-label BERT model.

3.4 Interpretability: Attention and Case Studies

We visualise attention weights from [CLS] to all input tokens (aggregated over heads and layers): an attention map (vertical axis = input tokens, heatmap = attention score; darker = stronger) and a bar plot of predicted probabilities for all eight classes (threshold line, ground-truth highlighted). We show two cases: (1) a complex multi-label case (five specialist referrals: Haematology, Nephrology, SOR, Cardiology, Pulmonology)—Fig. 5: attention distributed across renal, haematology, and symptom tokens; (2) a nephrology-only case (male, 60, high creatinine)—Fig. 6: attention concentrates on `CREATININE_Q9` and renal tokens, model outputs high probability only for Nephrology. This supports clinical coherence (XAI). We do not claim attention implies causality; it is a useful heuristic for audit.

Error analysis. Errors stem from (1) missing information (absent symptom or interview token → FN or FP) and (2) overlap of symptoms (dyspnoea and chest pain shared by Cardiology and Pulmonology). We distinguish data insufficiency (need more questions/lab values) from semantically justified confusion (syndrome overlap). Cardiology vs Pulmonology confusions often occur when both tokens are present; FN for Haematology sometimes when key symptom or bleeding-history tokens are absent (extending the question bank could reduce under-referral). Cases with low confidence or near-threshold predictions can be flagged for dual review or additional interview questions.

Case study. Example: input `AGE_60 SEX_M HGB_Q5...CREATININE_Q9 ALT_Q5` (male, 60, creatinine in band $Q9$). Model outputs high probability for Nephrology (e.g. 0.99), low for others. `CREATININE_Q9` drives the prediction; attention map (Fig. 6) highlights it. Elevated creatinine/urea are primary indicators for nephrology referral. Complex case: five referrals; input includes anaemia, renal, bleeding, infection, chest pain, dyspnoea tokens; model assigns high probability to all five; attention (Fig. 5) spread across renal, haematology, and symptom tokens—demonstrating multi-specialist routing in a single forward pass.

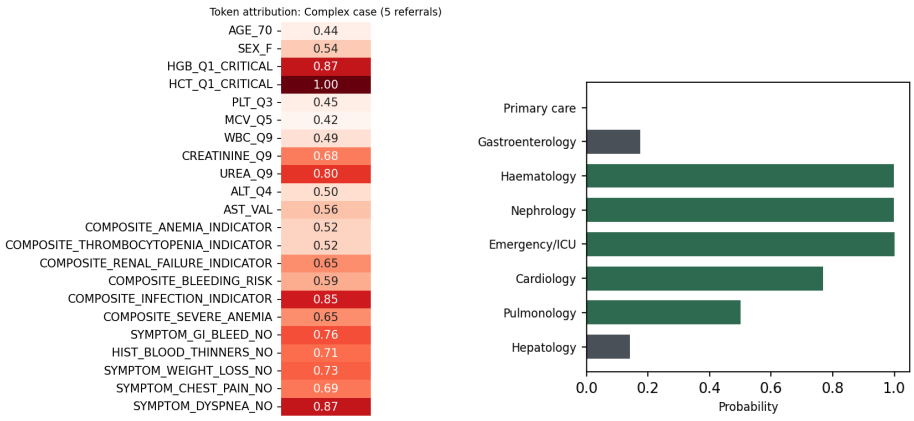

Fig. 5. Complex multi-label case: (a) attention map; (b) predicted probabilities (ground-truth highlighted).

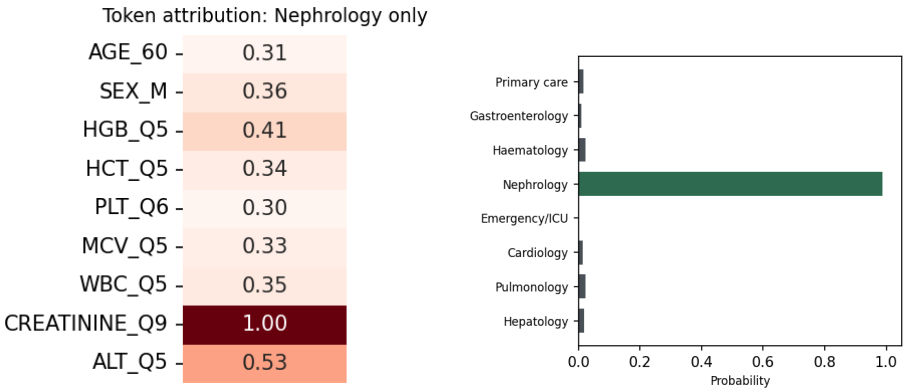

Fig. 6. Nephrology-only case: (a) attention map (creatinine $Q9$); (b) predicted probabilities (ground-truth highlighted).

4 Discussion and Conclusions

We presented a BERT-based model for multi-label clinical routing with limited patient history. The model is pre-trained on the hybrid corpus with MLM and fine-tuned for triage using quantised lab parameters and a minimal adaptive interview. The methodology includes a reproducible definition of each interview item, patient-level honest validation, and a cost-sensitive focal loss that reflects clinical priorities. Results confirm strong discriminatory power (high ROC AUC for SOR and critical classes), good calibration (low ECE), and interpretable attention patterns. Confusion between Haematology and Nephrology is consistent with shared lab patterns; in deployment, dual review or additional questions may be warranted when both are flagged. A small vocabulary ($\sim$140 tokens) keeps inference fast for real-time decision support in primary care.

Limitations. Synthea is synthetic and may be more regular than real data; hybridisation with MIMIC mitigates but does not eliminate overfitting risk. MIMIC is single-center and US-based. We have not performed prospective validation in primary care or on external multi-center datasets. The vocabulary is fixed at design time; adding new lab parameters requires re-building the tokeniser. The clinical cost matrix was set by expert opinion and was not learned from outcomes; its impact on referral quality remains to be validated. The ICD-to-specialist mapping is fixed and may not match local practice or coding conventions. The model does not capture temporal evolution across encounters (each sample is a single encounter); longitudinal triage and trajectory-based routing are out of scope and could be addressed in future work.

Contributions. (1) Multi-label BERT-based triage pipeline with eight referral classes, avoiding the single-label constraint ill-suited to comorbidity. (2) In-domain MLM pre-training on a hybrid corpus (Synthea + MIMIC) with cost-sensitive focal loss; reported ROC, precision–recall, ECE, confusion, and attention-based interpretability. (3) Patient-level honest validation and adaptive interview driven

by reference ranges and a question bank (auditable path from lab to routing). (4) High ROC AUC for SOR and good calibration, supporting deployment with clinician oversight.

Future work. We plan to (a) quantify the impact of the clinical cost matrix (cost-sensitive vs standard BCE); (b) extend the question bank from error analysis; (c) conduct prospective validation in primary care and on external datasets; (d) integrate with EHR and lab APIs for live deployment and optional A/B testing of the adaptive interview; (e) report stratified metrics by clinical subgroup (e.g. ICD chapter or referral pattern) to aid interpretation of aggregate scores, as suggested for heterogeneous hybrid corpora; (f) explore longitudinal encoders over multiple encounters.

Acknowledgements

The authors would like to thank Prof. Jarosław Wąs and MSc Filip Kamiński for their invaluable guidance, discussions, and support throughout this project. We also extend our deep gratitude to Aleksandra Hardzina for her expert medical consultations, which significantly contributed to the development of the clinical routing criteria and triage logic.

References

1. Z. C. Lipton et al., Learning to diagnose with LSTM recurrent neural networks. In ICLR, 2015.
2. A. Vaswani et al., Attention is all you need. In NeurIPS, 2017.
3. J. Devlin et al., BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT, 2019.
4. A. E. W. Johnson et al., MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 2023.
5. Synthea: Synthetic patient population. `https://synthea.mitre.org`