Minimal Data Maximum Impact: Lessons Learned from Real-World Unstructured Data in Paediatric Care

Jaskaran Singh Kawatra[1], Sebin Sabu[1], Pavithra Rajendran[1], Caroline Baumgartner[1], Avish Vijayaraghavan[1], Ewart Jonny Sheldon[1], John Booth[1], Neil Sebire[1], Shiren Patel[1], Alexandros Zenonos[2], Rebecca Pope[2]

[1]Great Ormond Street Hospital for Children NHS Foundation Trust
[2]Roche Products Ltd.

Abstract

Digital health records contains significant volume of pertinent, routine information locked within unstructured texts. Current processes requires costly human annotation from a limited number of expert annotators with sufficient domain knowledge and clinician's time for verification of the outcomes. Our proposed two-stage automated approach enables (1) training and validation of fine-tuned few-shot domain-specific models, firstly to retrieve relevant documents and then performing entity recognition on the retrieved document chunks for identifying correct span of texts based on the use case at hand and, (2) a "shadow deployment" pipeline testing an end-to-end solution in a pre-production environment. Our shadow deployment pipeline uses Large Language Models (LLMs) as an explainer-in-the-loop and Natural Language Inference (NLI) based verification approach to reduce the dependency on having a clinician to validate the outcomes of the solution. In this paper, we describe the experiments and results of deploying and testing our proposed approach within a real-world paediatric healthcare setting with a focus on histopathology reports of tumours, that can help answer clinical questions in a timely manner.

Introduction

National Health Service (NHS) England has published [1] several articles on the benefits of digitising health records with the help of Electronic Patient Record (EPR) systems. EPR systems provide a wealth of pertinent, structured information to aid clinical decision making. However, large volumes of potentially important routine data remain entirely wasted and locked within unstructured texts and thus unused as it is simply not feasible for manual extraction. By enabling an automated solution to extract useful information from unstructured texts, it can be used to guide clinical decision making based on past experience and information that is not currently available. Due to the sensitive nature of the data and strict information governance protocols within NHS Trusts, our proposed solution is focused on deploying custom and open-source based pipelines and models within our secure on-premise infrastructure.

[1]https://transform.england.nhs.uk/digitise-connect-transform/digitising-the-frontline/

Histopathology refers to tissue specimens examined for diagnosis including cancer and reports are written by clinicians based on the results. Our proposed solution [2] is evaluated on a specific use case related to extracting the final diagnosis from histopathology reports relevant to tumour condition. In a real-world setting, we find that a fairly generic clinical diagnosis code (e.g. *kidney tumour*) might be clinically coded but an unstructured histopathology report may provide a very specific diagnosis which is not coded in the EPR diagnosis codes (e.g. *Wilms tumour*, *intermediate risk*, *local pathological stage three*). Additionally, several factors important for treatment and clinical trial purposes may not be recorded in a structured manner such as the pathological stage of the tumour following surgery which is always commented by a clinician in the histopathology report and is rarely recorded in a structured manner in the patient notes elsewhere.

Our proposed solution (Fig. 1) is a two-stage approach and the key contributions are as follows:

1. The first stage, *training and validation* involves development of a modular, reusable end-to-end pipeline for experimenting, training and validating fine-tuning of few-shot domain-specific models to help automatically retrieving relevant documents and extracting useful span of texts within the retrieved document chunks. We show promising results with a limited amount of expert-labelled data and limited computation resources that does not require GPUs.
2. The second stage, *shadow deployment* for pilot testing within our secure on-premise pre-production environment and involves using lightweight, open-source Large Language Models (LLMs) as an explainer-in-the-loop and using Natural Language Inference (NLI) models for validation of the entity recognition predictions, thus aiming to reduce the time spent by a clinician for validating the extracted outcomes.
3. Our experiments and evaluation were carried out with minimal data consisting of 126 expert-annotated text chunks for retrieval classification and further tested on 300 expert-annotated text chunks with the best model performance achieving **0.897** F1-score. Similarly, minimal data consisting of 209 expert-annotated text chunks

[2]Source code will be open-sourced upon acceptance

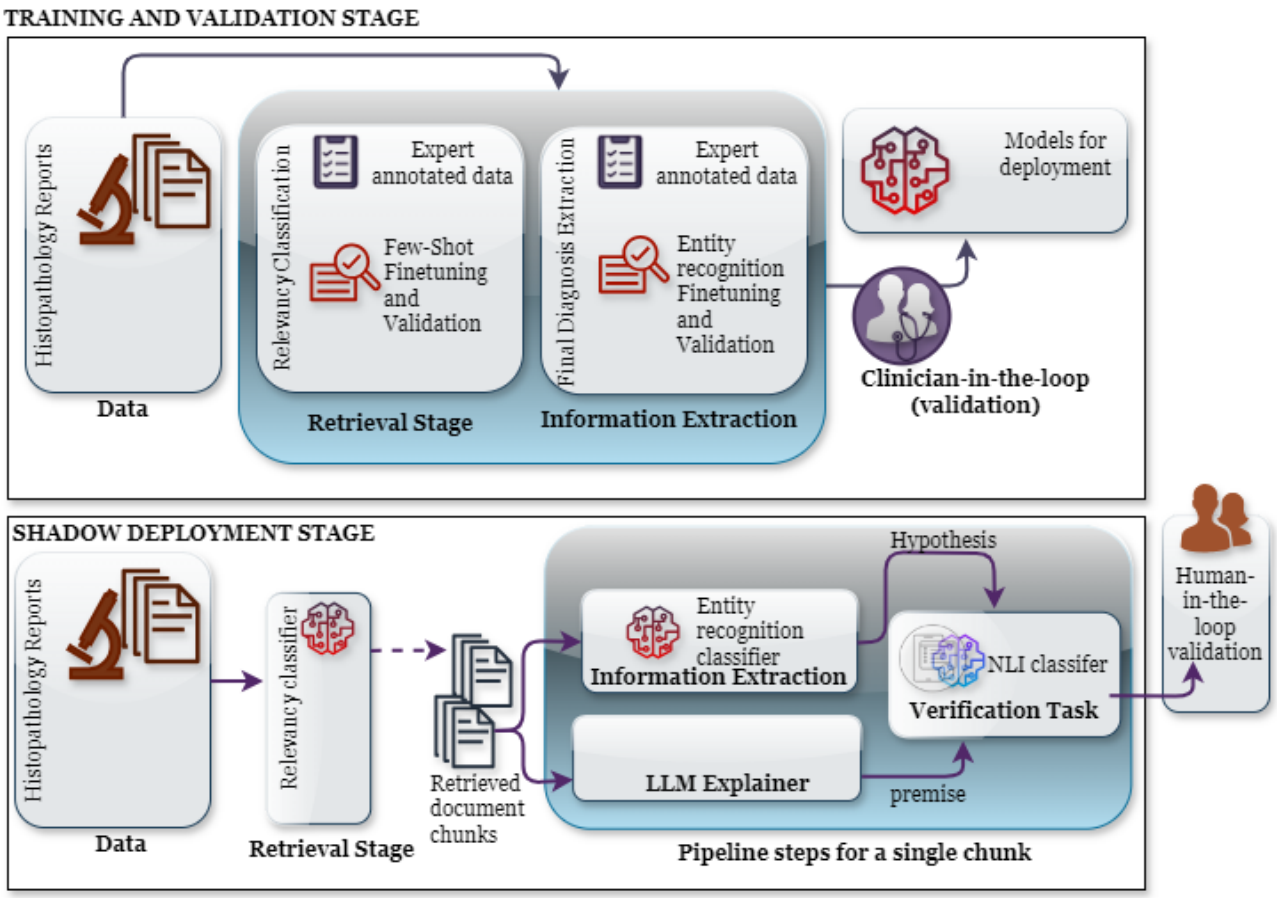

Figure 1: End-to-end architectures of our proposed approach depicting the two stages *training and validation* and *shadow deployment*.

REPORT 1

A. LYMPH NODE AORTA: NO EVIDENCE OF TUMOUR. B. RIGHT KIDNEY AND TUMOUR: POST CHEMOTHERAPY WILMS' TUMOUR; MIXED TYPE, INTERMEDIATE RISK, LOCAL PATHOLOGICAL STAGE 3 (DUE TO VIABLE TUMOUR AT RESECTION MARGIN). C. LYMPH NODE, INTERNAL ILIAC:NO EVIDENCE OF TUMOUR. D. LYMPH NODE, BIFURCATION OF AORTA: NO EVIDENCE OF TUMOUR. E. LYMPH NODE, RENAL ARTERY: NO EVIDENCE OF TUMOUR. F. LYMPH NODE, ANTERIOR AORTA: NO EVIDENCE OF TUMOUR. G. LYMPH NODE, SUPRAHILAR: NO EVIDENCE OF TUMOUR. Reported by<CLINICIAN_NAME>, Consultant Paediatric Pathologist.

REPORT 2

Second opinion from <CLINICIAN_NAME> The material has been examined by <CLINICIAN_NAME> from <HOSPITAL_NAME>. The lesion is a neoplasm with focal melanin pigmentation and immunoexpression of the markers HMB45, SOX-10 and S100 protein. FISH for EWSR1 gene rearrangement is negative. The proliferation index is low with no pleomorphism or necrosis. Overall, the appearances might represent a S100 protein-poor variant of a plexiform or cellular blue melanocytic naevus, although the cellularity is unusual and a dermatopathology opinion has now been requested. Please note that while this opinion is in accordance with the original report, the overall findings in this case are clinically unexpected since the original lesion was believed to be a vascular anomaly, and hence further clinical pathological correlation is required. For this reason an in basket message will be sent to the requesting clinician. Reported <CLINICIAN_NAME>, Consultant Paediatric Pathologist

Figure 2: Two dummy histopathology tumour report conclusion section written by an expert clinician.

were used for experimentation and evaluation of the final diagnosis extraction and further tested on 187 expert-annotated text chunks with the best model performance achieving **0.956** F1-score.

Our code has been open-sourced and is available for use on GitHub.

Related Work

Entity Recognition

Prior work (Yadav and Bethard 2019) provides an extensive survey on traditional NER systems that focused on sequence to sequence tagging. Recent approaches (Fu, Huang, and Liu 2021) focus on considering NER as a span prediction task and with recent advancement in LLMs (Chang et al. 2024), the focus has shifted to utilising instruction-tuning capabilities (Aw et al. 2023) of LLMs that are shown to outperform several LLM-based approaches (Mayhew et al. 2023). However, limitations of such work includes availability of large annotated data labelled with predefined labels which may not perform well with zero-shot predictions of unknown labels. In our proposed work, we leverage latest approaches such as GLiNER (Zaratiana et al. 2024) and NuNER (Bogdanov et al. 2024) that have tackled this problem with the help of using a Bidirectional Language Model and considering matching entity type embeddings with the corresponding textual span representations.

Large Language Models for summary generation

Recent work (Liu et al. 2024; Kirstein, Ruas, and Gipp 2024) have investigated on potential interest of human annotators favoring summaries generated by LLMs, which has also sparked interest within the medical community (Tang et al. 2023). Other approaches (Fu et al. 2024) have investigated on the potential capabilities of smaller and lightweight models for summary generation. Given the resource constraints within real-world hospital setting (such as no availability of GPUs), we adapt lightweight LLMs for summarising texts.

Information Extraction - Histopathology Reports

There are numerous biomedical articles (Mitchell et al. 2022; Abedian et al. 2021; Huang et al. 2024, 2023) using NLP for information extraction from histopathology reports. Most of the work focus on rule-based systems and training machine learning models with large annotated data. We differ from the above work by (1) using minimal annotated data, (2) fine-tuned lightweight entity recognition models that have zero-shot capabilities for detecting unseen entities and (3) using LLM-generated summaries and NLI for verification of the predictions.

Natural Language Inference

NLI systems ensure that the premise contains all the sufficient information for supporting a given hypothesis. It has been used for verification of question-answering systems (Chen, Choi, and Durrett 2021) and datasets such as FEVER (Thorne et al. 2018) were developed for fact verification and testing methods combining information extraction and NLI systems. Recent work (Sanyal et al. 2024) discusses on fine-tuned LLMs are better than human for complex reasoning in the context of open-source NLI datasets. In our work, we use a pretrained NLI model that was trained on the tasksource dataset (Sileo 2023) that was developed for harmonizing all the open-source NLP datasets and improve zero-shot classification.

Proposed Approach

Our proposed solution is setup for its practical utility in resource-constrained environments, minimal requirement of expert-annotated data and reducing the time spent by a clinician in the loop for validation of the outcomes.

In this paper, we focus on utilising our proposed solution for a specific use case namely retrieving the final diagnosis information from the main histopathology reports related to tumour condition as follows: (1) Identifying relevant histopathology text chunks that contain the main conclusion information about tumour diagnosis and (2) In the relevant text chunks, identify the final diagnosis relevant to the tumour. Figure. 1 shows the two-stage approach which is explained in the below sub-sections.

Stage 1: Training and Validation

For any given use case, the first step involves identifying documents or unstructured texts that are relevant. Information extracted from the relevant documents are further used for identifying spans of texts that are considered as relevant outcomes. The pipeline uses a SetFit-based approach (Tunstall et al. 2022) that fine-tunes an open-sourced pretrained Sentence-Transformers (Reimers and Gurevych 2019) model in a contrastive manner and the learned text representations are further used to train a text classification head. We benefit from this architecture's contrastive learning approach as it is useful in scenarios with minimal labeled training data. The second step involves fine-tuning an entity recognition model for identifying text spans within the retrieved relevant documents based on the use case requirement. Open-source model architectures exhibiting zero-shot capabilities for extracting any entity types are leveraged here.

For both the steps, a minimal expert-annotated dataset is used for fine-tuning the respective models. Further, another additional expert-annotated ground-truth dataset is used to validate the models and the clinician is provided the results for consideration of finalising the models.

Stage 2: Shadow Deployment

The purpose of this stage is to perform a pilot testing of a shadow deployment within a pre-production environment. The first step involves retrieving documents using the retriever model trained in the previous stage. The retrieved documents are then chunked and based on the use case, an entity recognition model trained in the previous stage is used for predicting the span of texts that are representative of the final outcomes.

In parallel, to reduce the human-in-the-loop evaluation process, this stage has an open-source, lightweight LLM as

an explainer-in-the-loop. Instead of having a human evaluator, the purpose of the LLM is to have the document chunk as an input and produce a corresponding summary based on the input with respect to the expected final outcome. This summary or explanation generated by the LLM and the final outcome predictions from the entity recognition classifier is then used as an input for the verification task. For the verification task, the generated explanation is considered as the premise and the predicted final outcome as the hypothesis. The verification task uses an open-source pre-trained NLI model that provides the output as entailment, contradiction or neutral. This is then provided to a clinician-in-the-loop for manually assessing the data classified as "contradiction" in the verification step.

It should be emphasized that the reason why LLMs of larger sizes (like GPT-4 and Llama 3.1) weren't used to simply extract the final diagnosis were mainly hardware infrastructure and data privacy based. Any processing of patient reports is done on internal infrastructure and the use of larger models even for zero-shot inference is beyond current infrastructure capabilities. Moreover, it is difficult to statistically estimate the performance of LLMs for practical clinical use cases where the goal is to minimize the occurrences of false negatives; instances where the model misses a cancer diagnosis when it was indeed present. The design choices in this pipeline are made for optimal extraction of diagnostic information while working within practical constraints.

Experiments and Results

Datasets and Model Training

Histopathology tumour reports contain useful structured information including the final diagnosis, specimen types and negative information such as *"no evidence of tumour"*. It has to be noted that specimen types represent the tissue but site could be anywhere depending on where the tumour is, which means, a single report can have multiple negative and positive information regarding the tumour (see Figure. 2). Often, the conclusion section within the report would help us in identifying whether the reports are relevant to that particular condition and further, the final diagnosis would always be represented within the same section. For the purpose of this study, we use the text chunks typed within the "conclusion" section and stored in the backend of our EPR system.

Table. 1 provides details on the training and inference data used for each of the individual components, also emphasizing on the realistic nature of having an expert clinician for annotating the data and thus utilising a very minimal ground-truth data.

Pipeline Stage	Training Data	Inference Data
Relevancy Classification	126	300
Final Diagnosis Extraction	209	187
Verification Task	-	187

Table 1: Training and Inference ground-truth dataset distribution for each component within the two stages shown in Figure. 1

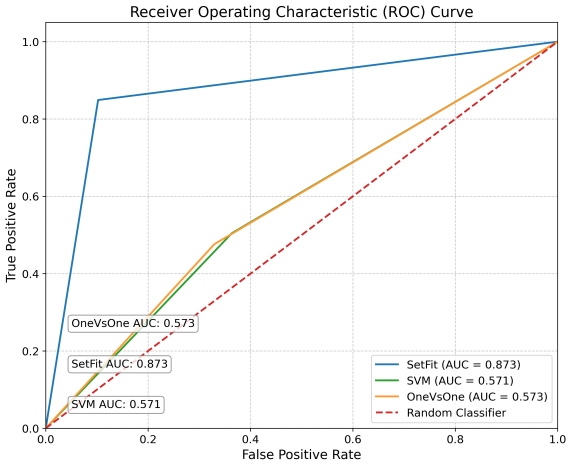

Figure 3: ROC Curves for models trained using Setfit-based approach, Linear SVM, and One-Vs-One classifier

Model Specifications We implemented and evaluated three distinct approaches for relevancy classification, with detailed specifications presented in Table 2. The primary model utilizes SetFit with a BioBERT backbone with a polynomial kernel. The hyperparameters were empirically determined: learning rate of 1.5215e-05 batch size of 16, and training conducted over 20 epochs with 35 iterations. For comparison, we developed two baseline models: a LinearSVC and a OneVsOne classifier with SVC, both employing TF-IDF vectorization limited to 5000 features. The LinearSVC was configured with C=0.1, while the OneVsOne implementation utilized an RBF kernel with C=1 and gamma set to scale.

For the extraction of final diagnoses, we fine-tuned two transformer-based models with complete specifications shown in Table 3. GLiNER-Bio-FT builds upon the base GLiNER model fine-tuned on PubMed data, while NuNER-Zero was initially trained on the NuNER v2.0 dataset. Both models underwent subsequent fine-tuning on 209 annotated histopathology records. The models share a DeBERTa-v3-large tokenizer and identifcal training configurations: primary learning rate of 5e-6 with weight decay of 0.01, secondary parameter learning rate of 1e-5, and linear learning rate scheduling with 0.1 warmup ratio. Training proceeded for 30 epochs using focal loss (/alpha=0.75, /gamma=2) with consistent batch sizes of 8 for both training and evaluation phases.

Training and Validation Stage The first stage involves training and validation of fine-tuning models for retrieving relevant text chunks containing the main conclusion specific to *tumour* condition within the histopathology reports and further, extracting text spans mentioning the final diagnosis within the retrieved text chunk using an entity recognition system respectively.

For the retrieval of relevant documents, a small subset of data was used for creating the ground-truth (see Table. 1) annotated with an expert clinician. Prior work (Arends et al.

Characteristic	SetFit Model	LinearSVC	OneVsOne Classifier
Base Model	SetFit with BioBERT backbone	LinearSVC	OneVsOne with SVC
Feature Extraction	-	TF-IDF	TF-IDF
Training Dataset	126 histopathology records	126 histopathology records	126 histopathology records
Hyperparameters			
Learning Rate	1.5215e-05	-	-
Batch Size	16	-	-
Epochs	20	-	-
Iterations	35	-	-
Seed	17	-	-
Kernel	poly	-	rbf
Max Iterations	200	-	-
TF-IDF Max Features	-	5000	5000
Classifier C	-	0.1	1
Classifier Gamma	-	-	scale

Table 2: Relevancy classification models specification

Characteristic	GLiNER-Bio-FT	NuNER-Zero
Base Model	GLiNER fine-tuned on PubMed	Trained on NuNER v2.0 dataset
Fine-tuning Dataset	209 histopathology records	209 histopathology records
Tokenizer	DeBERTa-v3-large	DeBERTa-v3-large
Training Parameters		
Learning Rate	5e-6	5e-6
Weight Decay	0.01	0.01
Others Learning Rate	1e-5	1e-5
Others Weight Decay	0.01	0.01
LR Scheduler Type	Linear	Linear
Warmup Ratio	0.1	0.1
Batch Size	8 (train and eval)	8 (train and eval)
Focal Loss Alpha	0.75	0.75
Focal Loss Gamma	2	2
Number of Epochs	30	30

Table 3: Final Diagnosis extraction models specification

2024; Beliveau et al. 2024) has shown the effectiveness in performance of SetFit-based models in data constrained settings for clinical data, motivating us to fine-tune the model with limited training data. We fine-tune an existing Sentence-Transformer model *BioBERT* [3] further fine-tuned for evidence extraction in medical claims (Deka, Jurek-Loughrey et al. 2022). After the initial fine-tuning of the sentence-transformer model, a One-vs-One SVM classifier is trained for the classification task. The model performance was validated on the inference data and further, results were validated by a clinician, which proved sufficient for the model to distinguish between relevant and non-relevant histopathology text chunks effectively. Table. 4 compares the performance of the fine-tuned Setfit-based model with a traditional SVM-based approach and a One-Vs-One SVM Classifier on the inference data; the best performance obtained using the Setfit-based approach. Figure 3 shows the ROC curve for the different models used for experimentation.

For the diagnosis identification step, we perform a holistic evaluation between five different models. First, we experiment using GLiNER, a state-of-the-art approach for NER that uses a bidirectional transformer as its backbone. It takes both the input text and entity type prompts as input. The model processes these inputs to generate token representations. For the input text, it uses a span representation layer to compute embeddings for potential entity spans. For the entity types, it uses a feedforward neural network to generate entity type embeddings. During training, the model is optimized to maximize the similarity scores between span embeddings and their corresponding entity type embeddings for correct entity spans, while minimizing scores for incorrect spans.

Next, we experiment with NuNER-Zero, another state-of-the-art approach that we use for evaluation which uses similar bidirectional tranformer architecture like the GLiNER approach but at the time of release showed the best compact zero-shot NER performance over GLiNER (`+3.1%` token-level F1-Score). The key difference between the two approaches is that unlike GLiNER, NuNER-Zero is a token classifier, which allows for the detection of arbitrarily long entities. We evaluate the two approaches with their

[3]We experimented with several sentence-tranformers model and this gave the best performance

Metric	SetFit	SVM	One-Vs-One Classifier
Accuracy	0.863	0.543	0.533
Precision	0.952	0.770	0.777
Recall	0.849	0.505	0.476
F1-Score	0.897	0.610	0.591

Table 4: Performance metrics comparison between Setfit-based approach, SVM and OneVsOne Classifier baselines

Model	F1 Score	Precision	Recall
GLiNER-Base	0.367	0.459	0.306
GLiNER-Bio	0.76	0.758	0.77
GLiNER-Bio-FT	**0.956**	**0.944**	**0.967**
NuNER-Base	0.46	0.34	0.708
NuNER-FT	0.92	0.902	0.937

Table 5: Entity Identification Performance Comparison

respective base models (**GLiNER-Base**, **NuNER-Base**), along with GLiNER's domain-specific model (**GLiNER-Bio**) trained with synthetic NER examples generated using an LLM model. We then assess the performance of these models after further fine-tuning on our in-house training data: **GLiNER-Bio-FT** (fine-tuned from **GLiNER-Bio**) and **NuNER-FT** (fine-tuned from **NuNER-Base**). Further, all these models were tested on our inference data (Table. 1).

Shadow Deployment stage This stage makes use of the fine-tuned models finalised in the *training and validation* stage. In this stage, our focus was to illustrate the potential capabilities of using existing LLMs and NLI models. We manually evaluated a random set of summaries generated by an open-source, lightweight LLM *Tiny-Llama-Chat* (Zhang et al. 2024) that ran efficiently on CPU servers. We found that the summaries were not prone to hallucination and hence we used the same model as an explainer-in-the-loop for generating the summaries of the text chunks. Further, for the verification step, the NLI model *deberta-base-long-nli* that has been fine-tuned on the open-source tasksource dataset (Sileo 2023) was used since it does not require any heavy annotation from our end and was also performing well.

Evaluation and Analysis

We evaluated the individual components within the two stages using their inference data predictions against the ground-truth expert curated data (see Table 1).

Training and Validation stage For the relevancy classification, the performance of the fine-tuned SetFit model, Linear SVM model and One-vs-One classifier model with an SVM head are evaluated on a set of 300 records consisting of 212 relevant histopathology records and 88 non relevant ones.

We observe that the Setfit-based approach significantly outperforms Linear SVM and One-Vs-One Classifier with SVM head (see Table 4) indicating strong utility in resource-constrained scenarios. This is also observed from Figure 3.

For diagnosis identification, the performance of several fine-tuned NER models was evaluated as discussed previously. Table 5 shows the results for classifying diagnosis entities correctly on our inference data. Our observations on the performance of the different models showed **GLiNER-Bio-FT** having the best performance followed by the **NuNER-FT** model. In general, we observed that, while models fine-tuned on NER data generated from PubMed articles like **GliNER-Bio** performed significantly better at zero-shot predictions in comparison to general models like **GliNER-Base** and **NuNER-Base**, it is not sufficient for specialised paediatric real-world data.

We also conducted a fine-grained analysis of model performance, specifically to identify negative diagnoses. Figure 5 compares the ability of each model to accurately capture negative diagnoses in text samples.

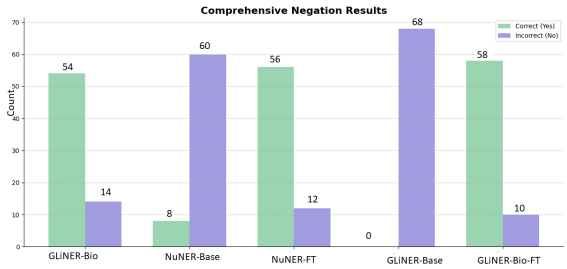

Figure 5: Comparison of negation handling performance across different Models

Across the board, all fine-tuned model versions were significantly performing better at identifying negative entities with **GLiNER-Bio-FT** identifying the most number of negative diagnoses correctly (58 out of 68).

Shadow Deployment stage The inference data and the predictions from the final diagnosis classification were used for the purpose of testing the pipelines within the shadow deployment stage. The *Tiny-Llama-Chat* generated a summary using the prompt: "*What is the final diagnosis in the report?*". The core of our evaluation in this stage focused assessing whether the diagnoses identified by *Tiny-Llama-Chat* are associated with the predictions we got from the final diagnosis extraction pipeline. The NLI model classifies the records across three different categories and assigns a confidence score for its predictions.

As an example, for the first report in Figure. 2, the following question was prompted to summarise based on the report content as shown below.

Question *What is the final main diagnosis from the report?*

LLM generated summary *The final main diagnosis from the report is "Wilms' tumour, mixed type, intermediate risk, local pathological stage 3 (due to viable tumour at resection margin)".*

Given the above generated summary which is considered as the premise and the predictions from the final diagnosis

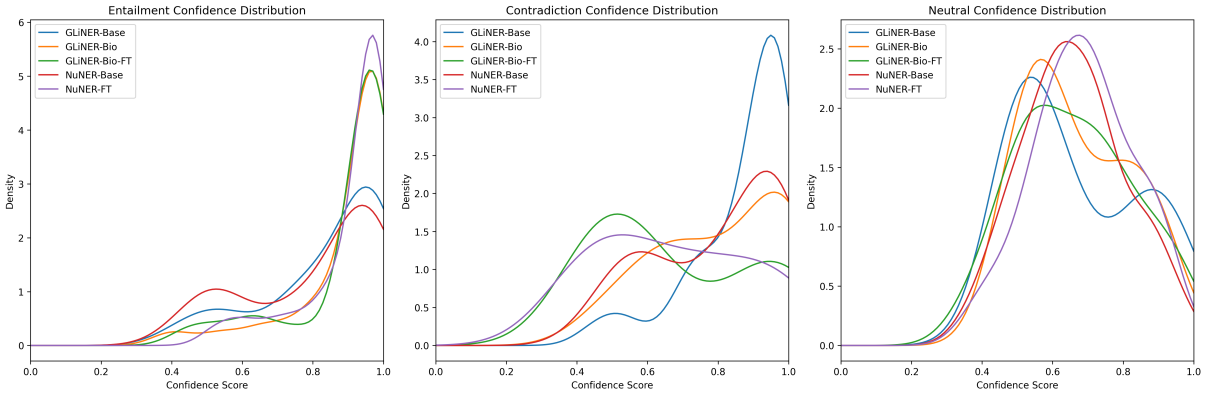

Figure 4: Kernel Density Estimates of NLI Confidence Score Distributions Across Each Category

classifier as the corresponding hypothesis, the NLI output is obtained as shown below.

Premise *The final main diagnosis from the report is "Wilms' tumour, mixed type, intermediate risk, local pathological stage 3 (due to viable tumour at resection margin)".*

Hypothesis *Wilms' tumour, mixed type, intermediate risk, local pathological stage 3*

NLI output *Entailment*

In Figure 4, the Kernel Density Estimates for each of the three categories namely *entailment*, *neutral* and *contradiction* with respect to the outcomes of the verification task pipeline which is further based on the outputs from different models experimented within the final diagnosis extraction pipeline.

We observe that for the Entailment Confidence Distribution (Figure. 4), all the diagnosis entity extraction models show high confidence, with peaks near 1.0. The premises generated by *Tiny-Llama-Chat* and the corresponding hypothesis using GLiNER approach based models have the highest and sharpest peaks, indicating very high confidence in entailment predictions. With NuNER models however, slightly lower and broader peaks are observed, suggesting more varied confidence in entailment. This is also reflected in the NLI classification results shown in Figure. 6 where the expected outcome is *entailment*.

For the Contradiction Confidence Distribution (Figure. 4), we observe that **GLiNER-Base** has the highest peak near 1.0, showing very high confidence in contradiction predictions. This should be expected as in the NER phase, **GLiNER-Base** had the lowest F1-score for identifying the diagnoses.

Limitations

Our pipeline has been developed as a modular, reusable code such that all the components can be replaced with other model architectures and models. One of the limitations of this work is the lack of computational resources such as GPUs for experimenting end-to-end solutions using LLMs.

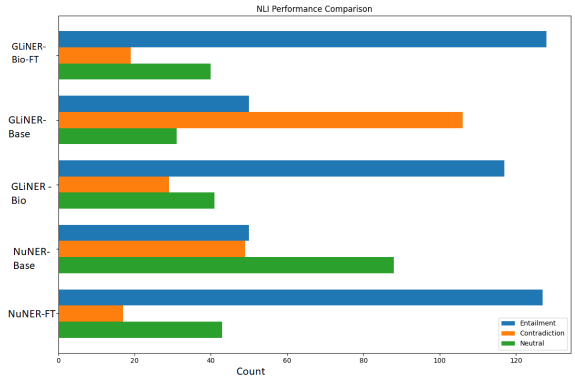

Figure 6: Total counts of the NLI-based classification based on the *Tiny-Llama-Chat* generated summaries and corresponding final diagnosis predictions. Results are reported for each individual final diagnosis extraction model.

We are also currently exploring post-monitoring the outcomes from the proposed approach, which at present, is not the focus of this paper.

Conclusion

Our proposed two-stage automated approach for retrieving relevant entity-based information from clinical texts is experimented and validated on histopathology reports to identify final diagnosis information within a real-world paediatric healthcare setting. Our results and analysis shows the potential value of this pipeline and we believe this approach is adaptable beyond tumour reports into a range of diseases which will subsequently allow the use of machine learning / artificial intelligence for some kind of clinical decisions support in the future. Our future plan is to pilot this work for prospective reports across a timeline and validate it with more clinical experts, which will then allow us to move it into a production-ready environment. This work presents a novel potential use case for extracting highly specific clinical information which can further be used in a hospital setting for various downstream tasks and also enables further

research directions within information extraction for paediatric care.

Acknowledgements

We would like to thank Daniel Key and Alex Eze for their support with infrastructural development.

This activity is part of a collaborative working agreement between Great Ormond Street Hospital NHS Foundation Trust and Roche Products Ltd. M-GB-00021122 | January 2025.

References

Abedian, S.; Sholle, E. T.; Adekkanattu, P. M.; Cusick, M. M.; Weiner, S. E.; Shoag, J. E.; Hu, J. C.; and Campion Jr, T. R. 2021. Automated extraction of tumor staging and diagnosis information from surgical pathology reports. *JCO clinical cancer informatics*, 5: 1054–1061.

Arends, B.; Vessies, M.; van Osch, D.; Teske, A.; van der Harst, P.; van Es, R.; and van Es, B. 2024. Diagnosis extraction from unstructured Dutch echocardiogram reports using span-and document-level characteristic classification. *arXiv preprint arXiv:2408.06930*.

Aw, K. L.; Montariol, S.; AlKhamissi, B.; Schrimpf, M.; and Bosselut, A. 2023. Instruction-tuning aligns llms to the human brain. In *First Conference on Language Modeling*.

Beliveau, V.; Kaas, H.; Prener, M.; Ladefoged, C.; Elliott, D.; Knudsen, G. M.; Pinborg, L. H.; and Ganz, M. 2024. Classification of Medical Text in Small and Imbalanced Datasets in a Non-English Language. In *Medical Imaging with Deep Learning*.

Bogdanov, S.; Constantin, A.; Bernard, T.; Crabbé, B.; and Bernard, E. 2024. NuNER: Entity Recognition Encoder Pre-training via LLM-Annotated Data. *arXiv preprint arXiv:2402.15343*.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3): 1–45.

Chen, J.; Choi, E.; and Durrett, G. 2021. Can NLI Models Verify QA Systems' Predictions? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3841–3854.

Deka, P.; Jurek-Loughrey, A.; et al. 2022. Evidence Extraction to Validate Medical Claims in Fake News Detection. In *International Conference on Health Information Science*, 3–15. Springer.

Fu, J.; Huang, X.; and Liu, P. 2021. SpanNER: Named entity re-/recognition as span prediction. *arXiv preprint arXiv:2106.00641*.

Fu, X.-Y.; Laskar, M. T. R.; Khasanova, E.; Chen, C.; and Tn, S. 2024. Tiny Titans: Can Smaller Large Language Models Punch Above Their Weight in the Real World for Meeting Summarization? In Yang, Y.; Davani, A.; Sil, A.; and Kumar, A., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, 387–394. Association for Computational Linguistics.

Huang, H.; Lim, F. X. Y.; Gu, G. T.; Han, M. J.; Fang, A. H. S.; San Chia, E. H.; Bei, E. Y. T.; Tham, S. Z.; Ho, H. S. S.; Yuen, J. S. P.; et al. 2023. Natural language processing in urology: automated extraction of clinical information from histopathology reports of uro-oncology procedures. *Heliyon*, 9(4).

Huang, J.; Yang, D. M.; Rong, R.; Nezafati, K.; Treager, C.; Chi, Z.; Wang, S.; Cheng, X.; Guo, Y.; Klesse, L. J.; et al. 2024. A critical assessment of using ChatGPT for extracting structured data from clinical notes. *npj Digital Medicine*, 7(1): 106.

Kirstein, F.; Ruas, T.; and Gipp, B. 2024. What's Wrong? Refining Meeting Summaries with LLM Feedback. *arXiv preprint arXiv:2407.11919*.

Liu, Y.; Shi, K.; He, K.; Ye, L.; Fabbri, A.; Liu, P.; Radev, D.; and Cohan, A. 2024. On Learning to Summarize with Large Language Models as References. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8647–8664. Association for Computational Linguistics.

Mayhew, S.; Blevins, T.; Liu, S.; Šuppa, M.; Gonen, H.; Imperial, J. M.; Karlsson, B. F.; Lin, P.; Ljubešić, N.; Miranda, L. J.; et al. 2023. Universal NER: A Gold-Standard Multilingual Named Entity Recognition Benchmark. *arXiv preprint arXiv:2311.09122*.

Mitchell, J. R.; Szepietowski, P.; Howard, R.; Reisman, P.; Jones, J. D.; Lewis, P.; Fridley, B. L.; and Rollison, D. E. 2022. A question-and-answer system to extract data from free-text oncological pathology reports (CancerBERT network): development study. *Journal of medical internet research*, 24(3): e27210.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sanyal, S.; Xiao, T.; Liu, J.; Wang, W.; and Ren, X. 2024. Minds versus Machines: Rethinking Entailment Verification with Language Models. *arXiv preprint arXiv:2402.03686*.

Sileo, D. 2023. tasksource: A Dataset Harmonization Framework for Streamlined NLP Multi-Task Learning and Evaluation. *LREC*.

Tang, L.; Sun, Z.; Idnay, B.; Nestor, J. G.; Soroush, A.; Elias, P. A.; Xu, Z.; Ding, Y.; Durrett, G.; Rousseau, J. F.; et al. 2023. Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, 6(1): 158.

Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 809–819.

Tunstall, L.; Reimers, N.; Jo, U. E. S.; Bates, L.; Korat, D.; Wasserblat, M.; and Pereg, O. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.

Yadav, V.; and Bethard, S. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.

Zaratiana, U.; Tomeh, N.; Holat, P.; and Charnois, T. 2024. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5364–5376. Association for Computational Linguistics.

Zhang, P.; Zeng, G.; Wang, T.; and Lu, W. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.