RWD31

Comparing Traditional and Large Language Models For Extracting Breast and Endometrial Related Clinical Features From Electronic Health Records

Keiran Tait [1], Joseph Cronin [1], Jamie Wallis [1], Robert Dürichen [1]

[1] Arcturis Data, Building One, Oxford Technology Park, Technology Drive, Kidlington, OX5 1GN UK
Keiran.tait@arcturisdata.co.uk

Introduction

- In oncology, pathological features such as MMR instability and ER/PR receptor status are critical for identifying patients for specific medications or clinical trials.
- These clinical features are often found exclusively in free-text reports such as pathology reports, making them difficult to access or analyse in real-world evidence studies.
- Large Language Models (LLMs) demonstrate promising potential for extracting oncology markers from unstructured real-world data (RWD) at scale.
- Nonetheless, concerns about accuracy and hallucinations (misinterpretations) remain when comparing LLMs to domain-specific Natural Language Processing (NLP) models.

Objectives

We have developed ArcTEX (Arcturis Text Enrichment and Extraction) model to support high-quality real-world evidence (RWE) studies by extracting oncology related features with high accuracy.

1. Compare ArcTEX to traditional NLP models (RoBERTa[1], BioBERT[2]) and general-purpose open-source LLMs (Llama2[3] and Llama3[4]) to extract oncology markers from unstructured real-world data (RWD) at scale.
2. Compare the impact of different training schemes and optimisation strategies, including zero-shot learning, few-shot learning, and finetuning, and prompt engineering.

Methods

Dataset & Annotation:

- 2,151 individual reports were taken from a wider dataset of 77,693 fully-anonymised free-text pathology reports provided by Oxford University Hospital (min-max number of words per report: 6-3213 words; mean number of words per report: 341.7)
- Annotations were performed for 18 clinical features:
  - Endometrial cancer: FIGO stage, grade, p53, MMR, MLH1, MSH2, MSH6, PMS2, myometrial invasion, and lymphovascular invasion
  - Breast cancer: HER2, ER, and PR
  - Additional features: TNM staging (T, N, and M stages and edition used), blast cell percentage
- In total: 3,568 manual annotations incl. absence of clinical feature and different score (e.g. HER2 → positive/negative/not performed; Figo → 2a/2b/3a…)

Baseline models*:

BERT baseline models:

- Performance was compared against a RoBERTa, BioBERT as question-answering model (stage 1) and a mpnet_v2[5] model for stage 2. This sequence is the same for the ArcTEX model (right) for direct comparability of results.

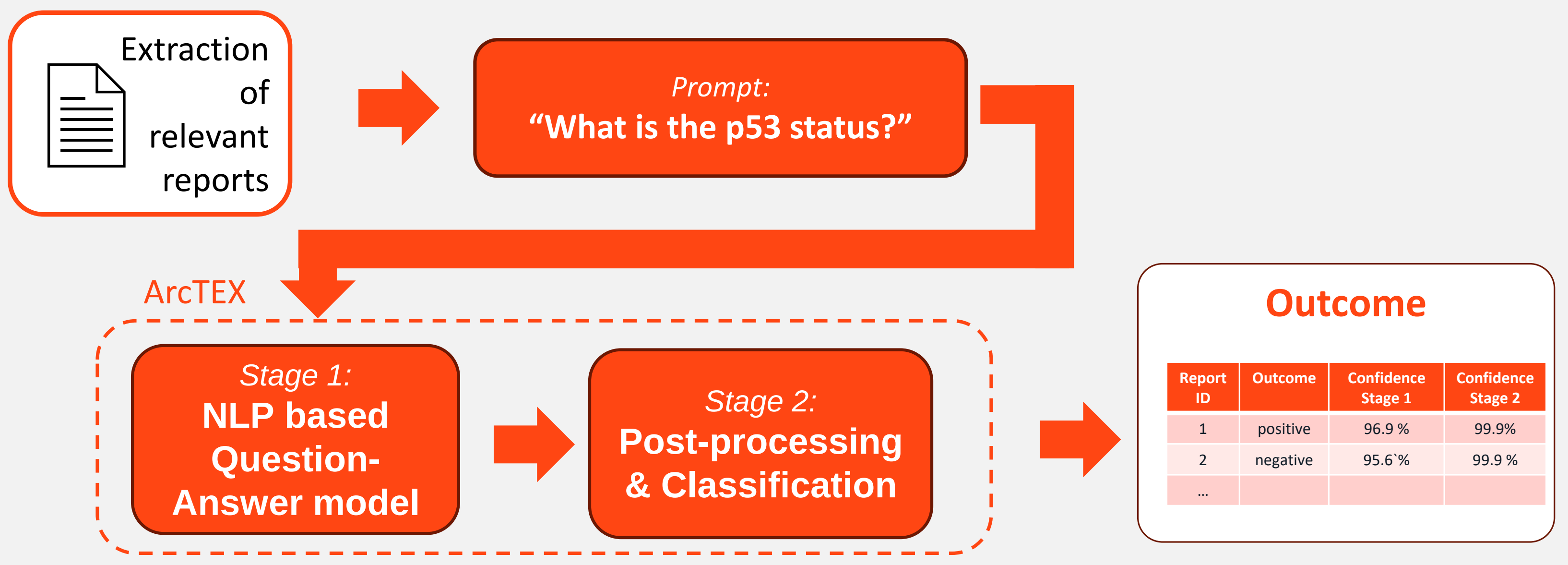
Large language models:

- Investigated LLMs: Llama-2-7B, Llama-2-7B-chat, Llama-3-8B, Llama-3-8B-Instruct
- Models were asked to provide the answer in a structured json (i.e.: {"HER2": "negative"}
- Multiple LLMs were optimised using the following techniques:
  - Prompt engineering (zero-shot, few-shot, role-based)
  - Unsupervised finetuning using *low rank adaptation (LORA)* using all available pathology reports[6]
- Example role-based prompt (*i.e. "You are a pathologist identifying particular biomarkers of interest. You are asked the question '{question}', regarding the following report: '{report}'. Please answer in the …")*

* All models accessed via HuggingFace platform

ArcTEX (Arcturis Text Enrichment and Extraction) model

- ArcTEX uses a two-stage process to extract and classify the results.
  - Stage 1: A finetuned BioBERT question-answering model extracts relevant text fragments from the report (i.e. "best described as wild-type")
  - Stage 2: A setfit classifier[7] and further post-processing classifies these into a set of predefined classes. This steps ensures that the outcomes of ArcTEX are always free of personal identifiable information by design.
- Each stage generates a confidence score, allowing the end user to threshold the output and analyse model performance in handling a given dataset.


Extraction of relevant reports
Prompt:
"What is the p53 status?"
ArcTEX
Stage 1:
NLP based Question-Answer model
Stage 2:
Post-processing & Classification
Outcome

Report ID	Outcome	Confidence Stage 1	Confidence Stage 2
1	positive	96.9 %	99.9%
2	negative	95.6`%	99.9 %
…			

Evaluation:

- Test set was composed of 100 reports per clinical feature, split between 50 containing the feature and 50 without (n = 1800).
- The training set was composed of all remaining annotated reports with no overlap of the test set (n = 4714, average = 261.9 reports per marker). Reports, both with and without annotated clinical features were used for training and testing.

Results

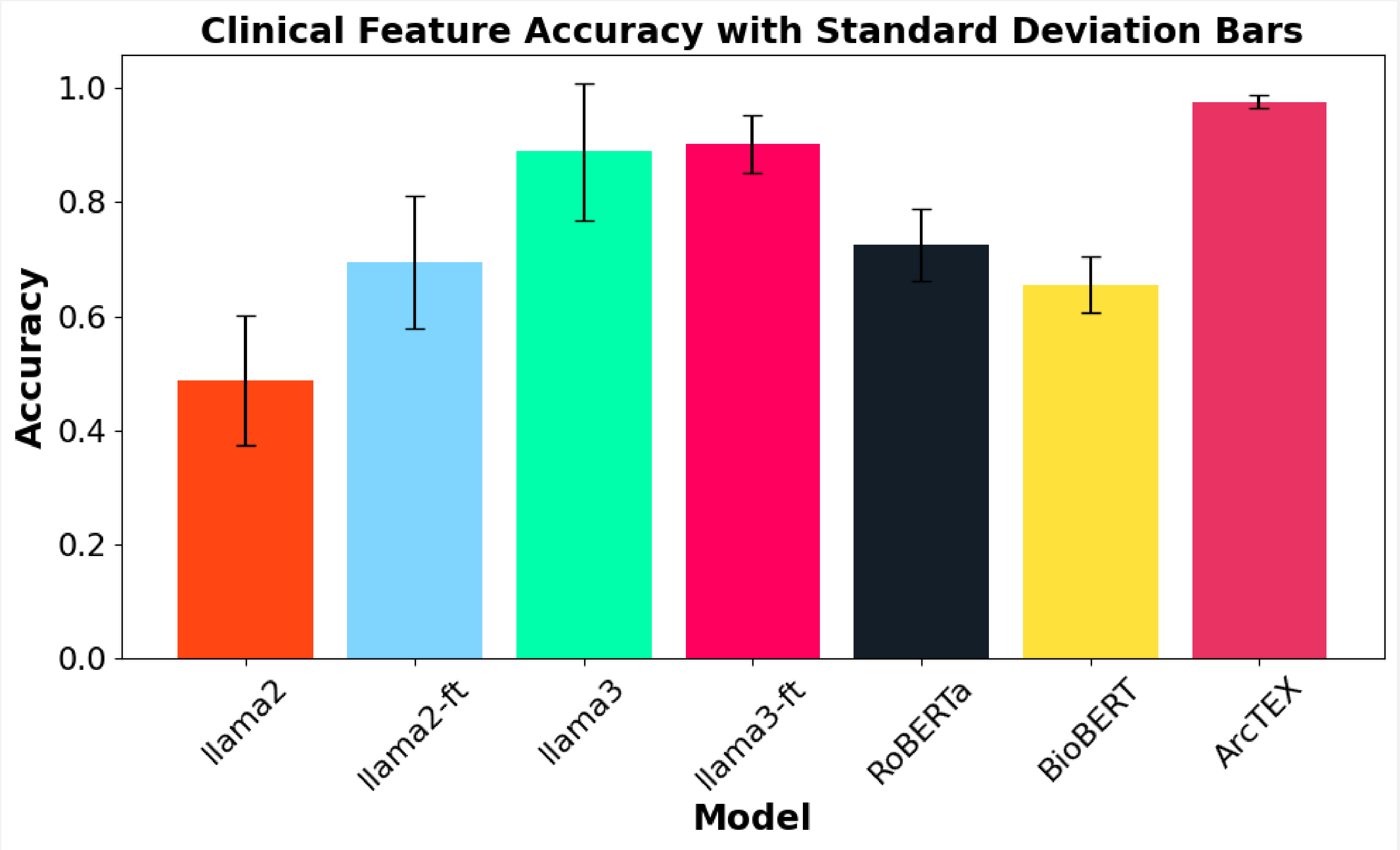


Figure shows results of specific LLMS and BERT based models. Llama labels have been shortened in figure and refer to:

- LLama2: Llama-2-7B-chat, zero-shoot, no further optimisation
- LLama2-ft: Llama-2-7B-chat, role-based prompt and unsupervised finetuning
- LLama3: Llama-3-8B-instruct, zero-shoot, no further optimisation
- LLama3-ft: Llama-3-8B-chat, role-based prompt and unsupervised finetuning

All other LLMs and optimisation techniques were tested but resulted in lower accuracy compared to the best LLM models (results not shown here).

- ArcTEX demonstrates superior mean accuracy (98.66%) and lower variation (standard deviation = 1.1%) in comparison to the other models.
- The best-performing open-source LLM was Llama-3-ft, a Llama-3-8B-Instruct with role-based prompts and LORA finetuning, achieving a mean accuracy of 90.23% (standard deviation= 5.1%) across clinical markers
- The remaining models scored progressively lower scores with RoBERTa (mean: 72.44% / std: 9.70 %), Llama2 (mean 69.54% / std: 11.6%), and BioBERT (mean: 67.67 % / std: 5.71%); all scoring less than 75% accuracy.
- Impact of optimisation of Llama3 model is low (comparison: llama3 vs. llama3-ft)
- Role-based prompting is superior compared to few-show learning for all Llama models

Conclusions

- ArcTEX demonstrates superior accuracy and consistency in extracting clinical features from pathology reports, outperforming both BERT-based models and LLMs, even after fine-tuning.
- Extensive fine-tuning is required for LLMs to match the accuracy of domain-specific models (in particular for Llama2); zero- or few-shot prompting remains insufficient.
- Untrained LLMs often generate incorrect output formats, complicating result interpretation.
- Unlike LLMs, ArcTEX also provides confidence scores at each extraction step, offering deeper insights into how the extracted data can be effectively utilised.
- Both LLMs underperform compared to ArcTEX, highlighting the power of a model which is computationally less expensive, and require less finetuning to achieve the correct outputs. Furthermore, ArcTEX demonstrates superior accuracy with reduced deviation between clinical features compared to any comparable model.

Acknowledgements

The Real World Data Network has received research database ethical approval from the NHS Health Research Authority Yorkshire & The Humber - Leeds East Research Ethics Committee (REC Reference: 24/YH/0164). This work uses data provided by patients and collected by the NHS as part of their care and support. We believe using patient data is vital to improve health and care for everyone and would, thus, like to thank all those involved for their contribution. A pseudonymised dataset for this study was provided by the Oxford University Hospital Trust.

References

[1] Liu, Y. et al.: "RoBERTa: A Robustly Optimized BERT Pretraining Approach", arXiv:1907.11692, 2019

[2] Lee, J. et al.: "BioBERT: a pre-trained biomedical language representation model for biomedical text mining", arXiv: 1901.08746, 2019

[3] Touvron, H. et al.: "Llama 2: Open Foundation and Fine-Tuned Chat Models". arXiv: 2307.09288, 2023

[4] Dubey, A. et al.: "The Llama 3 Herd of Models", arXiv: 2407.21783, 2024

[5] Song, K. et al.: "MPNet: Masked and Permuted Pre-training for Language Understanding", arXiv: 2004.09297, 2020

[6] Hu, E. J. et al.: "LoRA: Low-Rank Adaptation of Large Language Models", arXiv: 2106.09685, 2021

[7] Tunstall, L. et al.: "Efficient Few-Shot Learning Without Prompts", arXiv: 2209:11055, 2022