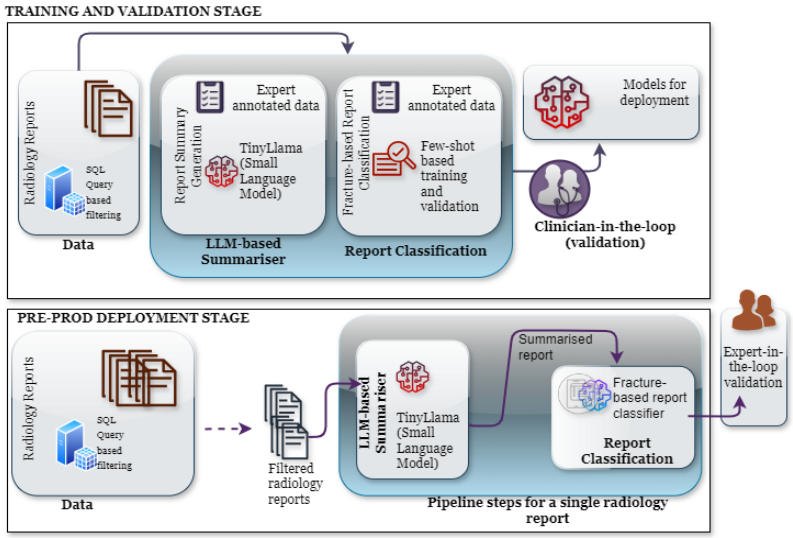
Generate-then-Classify: Can Large Language Models generate better summaries of radiology reports for automated Report Classification?

Sebin Sabu, Caroline Baumgartner, Pavithra Rajendran, Lucy Turriff, Nuwanthi Yapa Mahathanthila, Alexander Chesover, Shiren Patel
Great Ormond Street Hospital NHS Foundation Trust

BACKGROUND

- **In 2020,** the Patient Safety Team at GOSH noted that there had been a notable in incidents relating to pathological fractures – a review was completed of 18 incidents, between Aug 19 and May 21. There have been subsequent incident reports relating to fractures since the initial review was carried out, and it is anticipated there may be other cases not reported or identified through the incident reporting process at the Trust. Therefore, a QI project was established.
- **Quality Improvement(QI)** team is currently working on a manual process for identifying the pathological fractures present in radiology reports as part of a project to **'Reduce the number of fractures for patients at risk of fractures at GOSH'**
- Outcome measure of the QI project is the overall number of inpatient fractures at increased risk of fracture, the team recognize that it might take some time to see a statistical change for this.
- One of the measures is to identify preventable fractures from radiology reports. The NLP team at DRIVE help them to make this process easier by using a type of AI, known as **Natural Language Processing (NLP) techniques** to build **GOSH-specific AI models** for **identifying the occurrence of pathological fractures from radiology reports.**.



METHODOLOGY

Trial 1

setfit

- ✓ Custom trained AI models using radiology reports, expert-annotated labels and using SetFit, an AI-based approach.
- ✓ Trained AI model fails to understand the complex contextual meaning of words despite the expert annotated labels.
- ✓ Performance accuracy is 52%

Trial 2

setfit



- ✓ "Small" Large Language Model (LLM) *TinyLlama* is used to generate a summarised version of radiology report in context with pathological fracture information.
- ✓ Trained AI model using the summarised data and expert annotated labels.
- ✓ Performance accuracy is improved to 75%

Trial 3

setfit



- ✓ Following the same process as Trial 2 with oversampling the data since the occurrence of "pathological fractures" in radiology reports is very small.
- ✓ Performance accuracy is improved to 92%

RESULTS

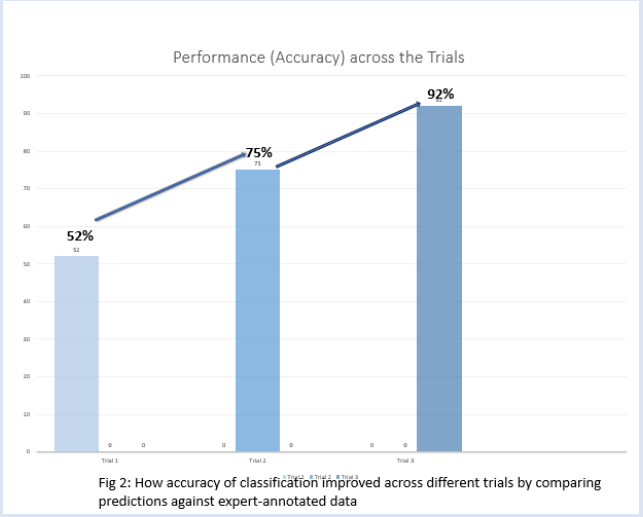


Fig 2: How accuracy of classification improved across different trials by comparing predictions against expert-annotated data

High Quality Expert Annotation enhances AI

Identifying the occurrence of a "pathological fracture" in the radiology reports is a complex process; it requires expertise understanding of the text within the report

❑ **Expert annotated radiology reports collected by the QI team is used for developing GOSH specific AI models.**

CONCLUSION

- **Summaries generated using Large Language Models can significantly improve automatically prediction of radiology reports.**
- **Our custom-trained GOSH specific AI models make use of very small labelled expert-annotated data showing the potential capabilities of using Natural Language Processing and adapting to usage in complex clinical text scenarios for Quality Analysis and adaptable for similar use cases.**
- **Significant savings in time reducing the processing of radiology reports from several days if done manually to a few hours using our own in-house infrastructure.**