

# LLM-based Detection of Voice Recognition Errors in Radiology Reports: Development, Validation and Impact of Personalised Feedback

Dr Adam Rushford, Dr Jaymin Patel, Dr Robert Dunk, Dr Rebecca Greenhalgh  
Hexarad Group, London, United Kingdom



## Background

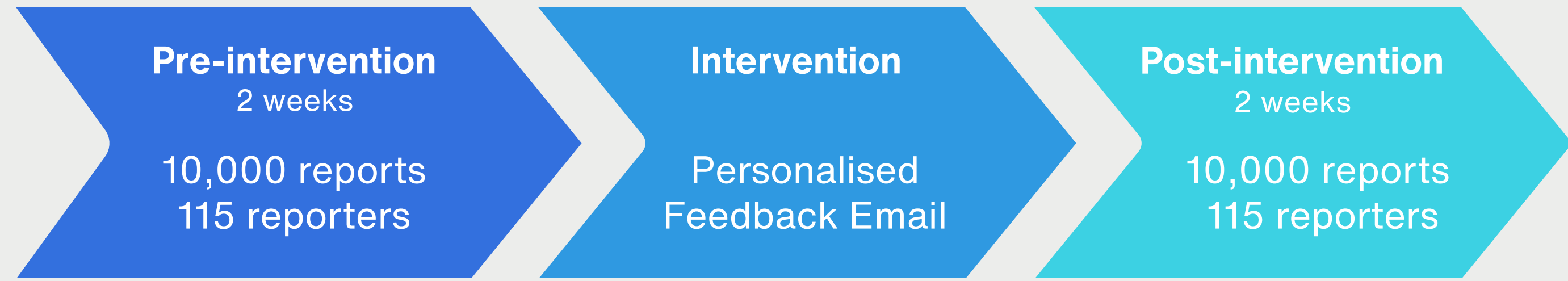
Voice recognition (VR) software is widely used by radiologists and reporting radiographers but remains susceptible to errors that compromise report quality, interpretability and reader confidence. Published estimates suggest VR errors occur in 6-44% of reports, depending on definitions and methodology<sup>1-3</sup>. We developed a large language model (LLM)-based tool to detect VR errors at scale and used the output to deliver personalised feedback to reporters, aiming to reduce error rates. To our knowledge, this is the first deployment of an LLM-based VR error detection and individualised feedback system in clinical radiology practice.

## Methods

VR errors were defined as text anomalies degrading report quality beyond simple grammar or spelling mistakes. Broadly, there were categorised as:

| Error Type            | Example                                       |
|-----------------------|-----------------------------------------------|
| Nonsense text         | "proximal tibial unicorn spur"                |
| Extraneous statements | "the right lung is clear hi Mark how are you" |
| Omitted words         | "The uterus is an axial position"             |
| Additional words      | "Intact the glenoid labrum"                   |
| Duplicated text       | "and and"; "the the"                          |
| Homonym errors        | "opacity in the write lung"                   |

We used Claude Sonnet 4.5 as the base LLM. The prompt was refined iteratively against a development set of 150 randomly sampled reports, using chain-of-thought reasoning and few-shot learning. Heuristic post-processing reclassified non-erroneous phrases predictably miscategorised by the model, such as common anatomical abbreviations. Performance was evaluated on an independent reference standard of 600 manually labelled reports.



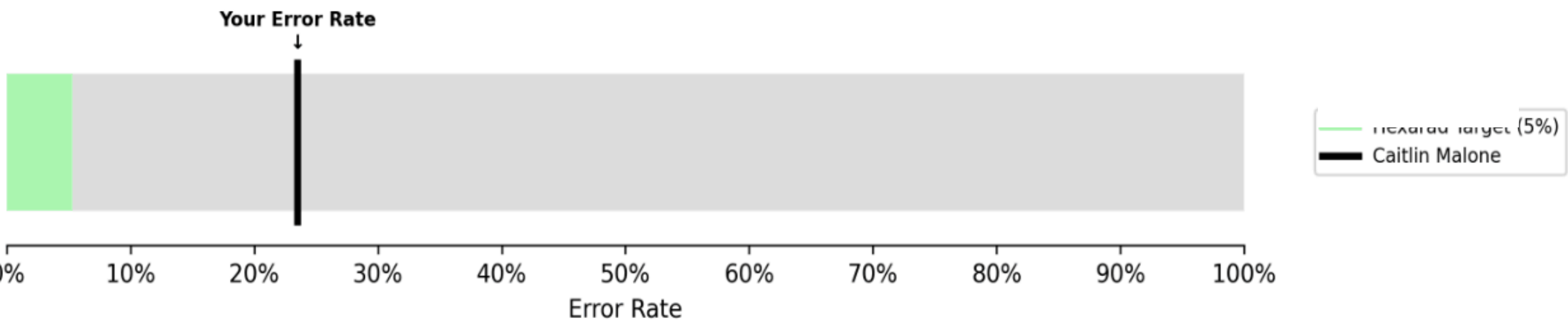
Analysis was restricted to reporters active in both periods, averaging at least 20 studies per week, with pre- and post-intervention periods matched by modality mix within the same reporters. Overall error proportions were compared using a chi-squared test.

### Your Personalised Feedback

Between January 19th and February 22nd, you reported **230** studies.

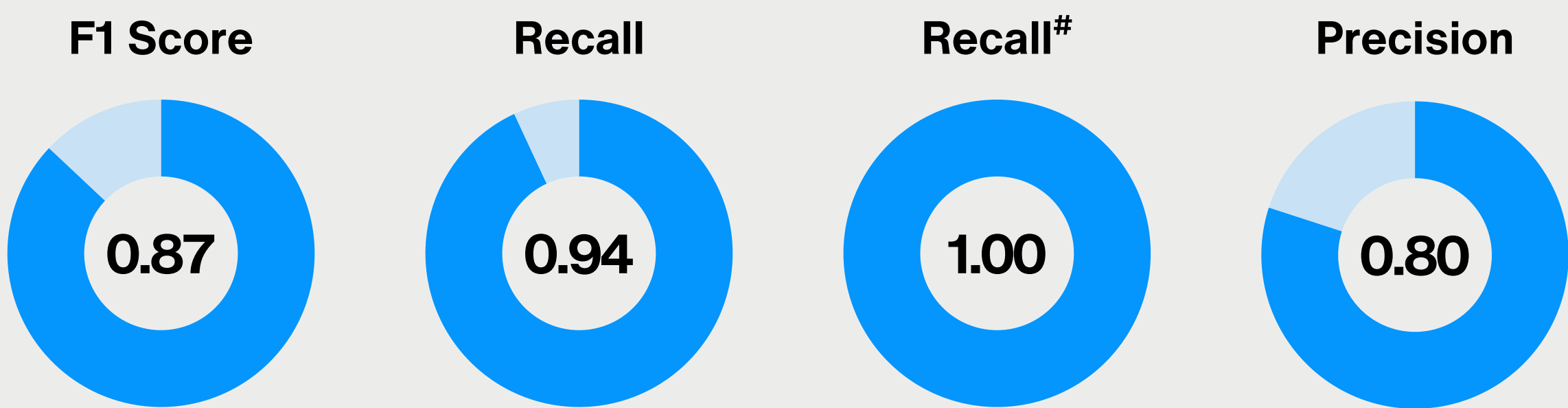
- Of the **230** studies you reported **54 reports** contained one or more errors
- This means that **23.5%** of your reports contained one or more error

We want to see a target of less than 5% of reports containing an error.



## Results

### Model Performance

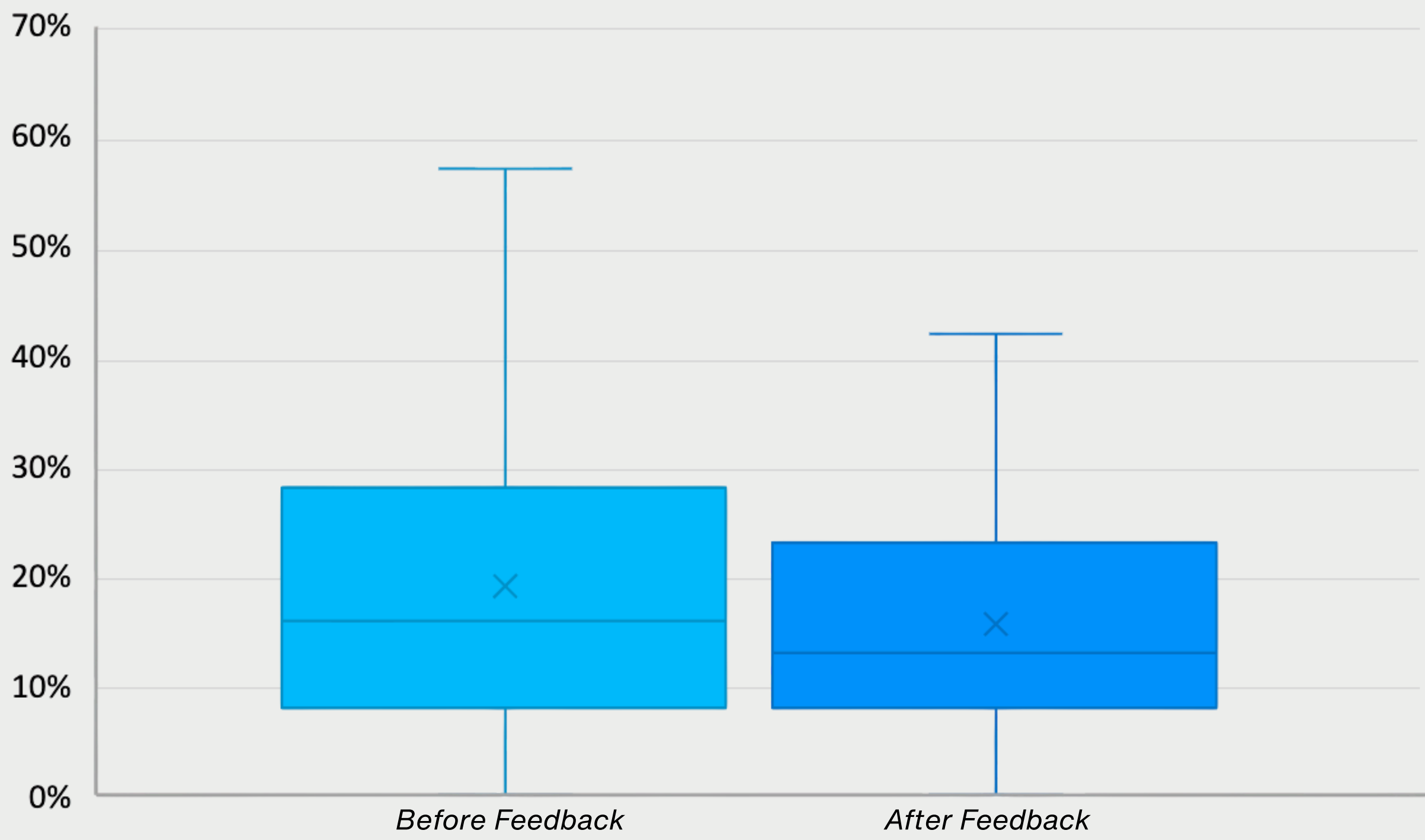


# The model achieved a recall of 100% for clinically significant VR errors, defined as VR errors that would likely require an addendum

### 12.3% Reduction in VR errors

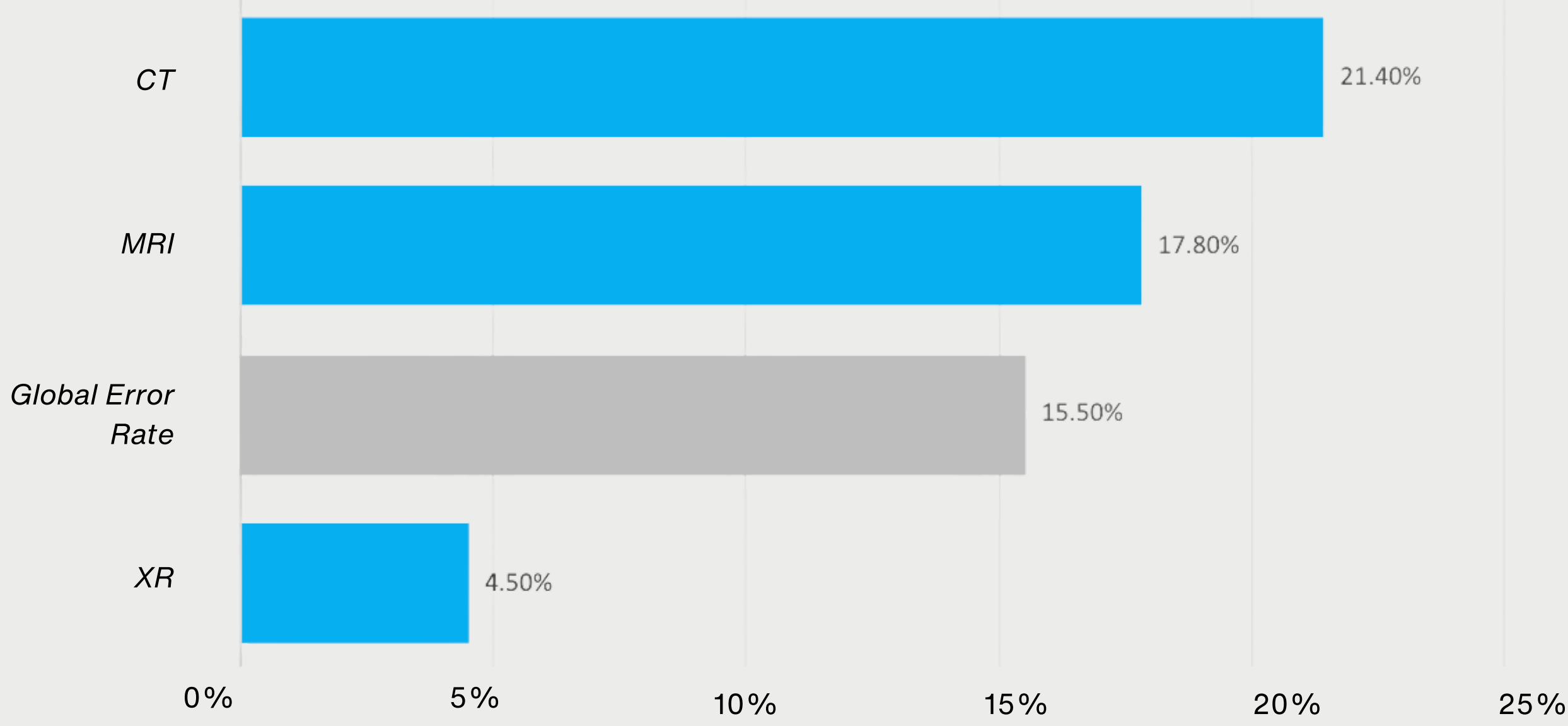
VR errors were identified in 1,504 of 10,000 pre-intervention reports (15.0%) and 1,319 of 10,000 post-intervention reports (13.2%), denoting a 12.3% reduction in VR errors ( $\chi^2 = 13.96$ ,  $p = 0.0002$ ) following the provision of LLM-generated personalised feedback.

### VR Error Rates (pre and post-intervention)



At baseline, VR errors were most common amongst CT and MRI:

### VR Error Rates (by modality)



## Conclusion

- LLMs can detect VR errors in radiology reports effectively at scale
- Personalised feedback produced a statistically significant reduction in error rates, resulting in the improves quality of radiology reports
- Combining LLM-enabled QA with individualised feedback shows promise for continuous professional development

1. Schmidt et al. Radiol Artif Intell. 2024, 6(2).  
2. Chang et al. J Digit Imaging. 2010, 24(4), 724-728  
3. Quint et al. J Am Coll Radiol. 2008, 5(12), 1196-1199

### Contact

