

Pilot Evaluation of a Custom GPT Model for Automated Quality Assurance of PET-CT reporting

Jack Sheppard 1, Craig George-McDowall 1, Dafydd Ap Emyr 1, Alex Powles 1
1 Swansea Bay University Health Board (SBUHB)

AI-assisted QA pilot · PET-CT radiology reporting



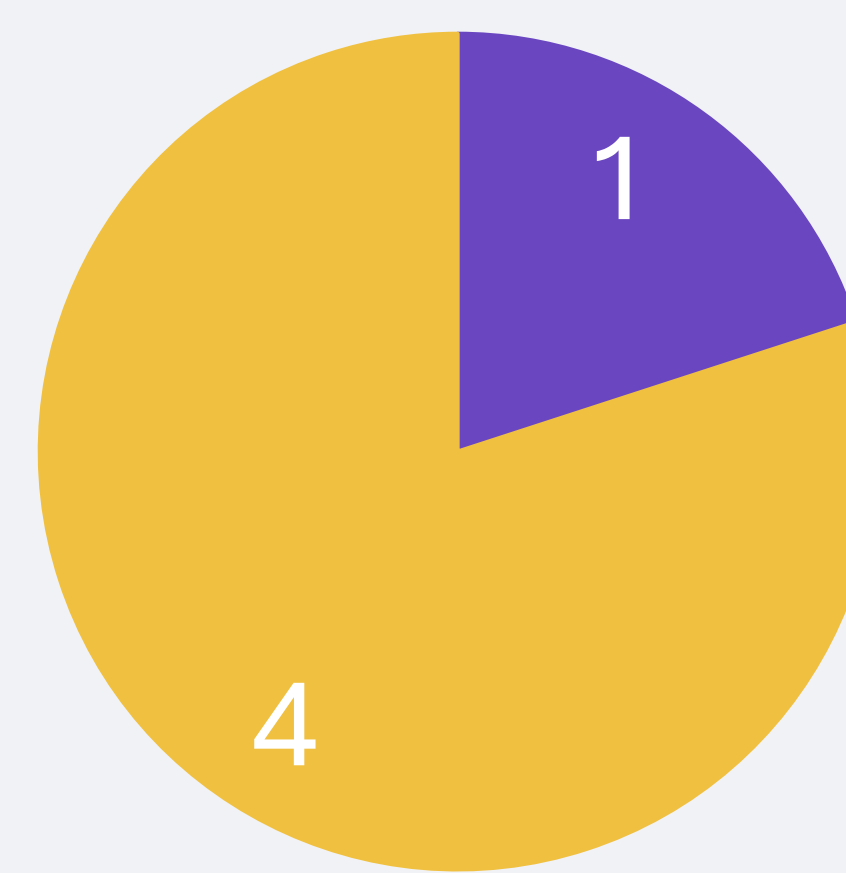
Scan QR code 1 for prompt, framework & example outputs



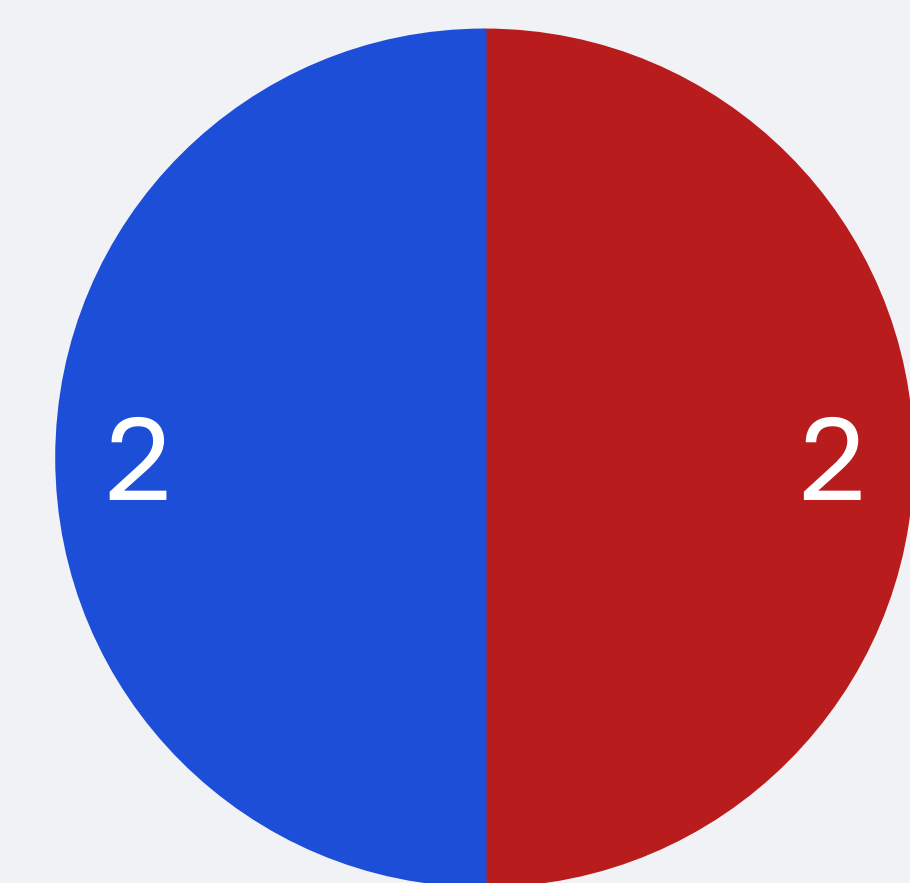
Scan QR code 2 for more detail on the AI-detected clinically significant errors

AI-detected clinically significant errors

- True positive
- False positives



Potential implications of false positive errors without clinician oversight



- Negative
- Positive

AI Quality Assurance in PET-CT Reporting

1 Why This Matters

PET-CT reports are complex and high-stakes. Laterality, consistency and staging errors can affect patient outcomes, yet traditional double-reporting QA is unsustainable at scale. LLMs offer a scalable, automated solution for continuous report quality monitoring.

2 What We Did

A custom GPT was configured as a radiology QA tool and applied to 100 anonymised PET-CT reports. It used:

- a structured prompt
- an error detection framework
- a clarity refinement framework

Each report yielded two outputs:

- clinically significant errors** (laterality error, anatomical mislabeling, numerical contradictions, or major logical contradictions)
- non-clinical refinements**

3 What We Found

- 5 clinically significant errors flagged by AI
- 1 confirmed:** lung-lobe laterality error (findings vs. conclusion)
- 4 false positives (dictation artefacts / orientation)
- 276 non-clinical refinements generated (median 3/report).
- Processing: <1 sec/report.

AI Error Detection

AI flagged *5 clinically significant errors*. Human adjudication confirmed:

- True clinical error: 1**
- False positives: 4**

The true error was a nodule location lung-lobe laterality discrepancy between the findings and conclusion sections. False positives arose from dictation artefacts and radiology vernacular.

Discussion points

There were *276 clarity improvements* across 100 reports.

Median: 3 per report. Types:

- spelling corrections
- terminology standardisation
- sentence simplification
- improved conclusion clarity

All AI refinements were judged by radiologists to be meaning-preserving.

Conclusion

This pilot shows how clinician-designed AI can act as a continuous background QA layer, without replacing the radiologist, focusing on tasks prone to fatigue-related error:

- maintaining internal consistency
- detecting contradictions
- standardising terminology

THE BOTTOM LINE: AI-assisted QA may improve report clarity, consistency and safety while preserving full radiologist oversight.

KEY FINDING:

AI-assisted QA identified 1 clinically relevant laterality error and 276 refinements across 100 PET-CT reports