Evaluation of Large Language Models with clinical guidance for vetting outpatient MRI lumbar spine referrals

William Clackett[1], Hatim Alsusa[1], Hannah Watson[2], Antanas Kascenas[2], David Scott[1], Avinash K. Kanodia[1], T. Barry Oliver[1], and Alison Q. O'Neil[2]

[1]Department of Radiology, NHS Tayside, Dundee, United Kingdom

[2]Canon Medical Research Europe, Edinburgh, United Kingdom

Introduction

Vetting of radiology referrals ensures imaging requests are clinically justified and appropriate for the diagnostic question. This process results in assigning an outcome for the patient referral; either rejecting the request or accepting the request (with an urgency assigned).

Rising annual demand for MRI examinations is exceeding radiologist recruitment rates and is leading to an increased vetting workload [1,2]. This rising vetting workload may lead to greater vetting errors, which can result in unnecessary cost as well as increased patient risk. Large Language Models (LLMs) may offer a promising solution to accelerate vetting whilst retaining sufficient vetting accuracy. These models have shown success in extracting concepts from radiology documents without requiring extensive clinical training data [3, 4]. In concept learning appears to be non-inferior to pre-training offering a solution to the significant computational resources required when pre-training models.

This study evaluates the role of LLMs for vetting MRI lumbar spine referrals from GPs for the problem of low back pain. At NHS Tayside, the number of referrals of these studies from GPs have increased from 1500 to 2500 annually over 10 years. This type of referral is an excellent candidate for exploring the role LLMs can play in vetting as the vetting decision is largely based on the text-based referral only.

We aim to test state-of-the-art LLMs in the vetting of synthetically created GP-requested MRI studies.

Methods

Synthetic MRI lumbar spine referrals were created by radiology registrars from a GP's perspective for patients with low back pain. The referrals were intentionally diverse in length, acronym usage, detail level, and clinical appropriateness, including patient age and sex information.

Vetting was performed by two board-certified radiologists (one musculoskeletal, one neuroradiologist) with 11 and 18 years of experience respectively. They categorised referrals as: (a) Accept - Routine, (b) Accept - Urgent, or (c) Reject [reason for rejection]. A separate 'Safety' column recorded possible MRI contraindications within the referral. Disagreements were resolved by a separate senior radiologist (musculoskeletal) with 26 years' experience.

We tested three off-the-shelf large language models (GPT4, Claude Opus, Google Gemini), implemented with a task-appropriate prompt. The prompt included instructions for vetting outcome assignment and MRI contraindication identification. We included separate prompts with varying degrees of 'context'; this context was the NHS Tayside referral guidance (vetting task), MRI safety guidance (MRI contraindication task) and example referrals with correct vetting outcome and safety outcome.

The final test set included 120 referrals. Inter-rater agreement between radiologists was measured using Cohen's Kappa. Performance was assessed using accuracy for the vetting task and recall, precision and F1 scores for contraindication identification.

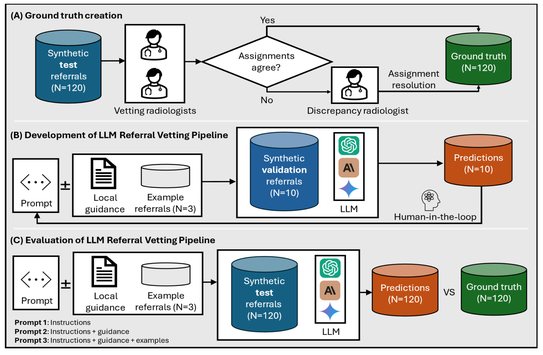


Figure 1: *Study protocol.*

Results

Vetting Assignment	Radiologist 1	Radiologist 2	Cohen's Kappa	Ground truth
Vetting outcome				
Accept - Routine	69	69		70
Accept - Urgent	23	26	0.76	22
Reject	28	25		28
MRI contra-indications				
Present	8	9	0.68	9
Absent	112	111		111

Table 1: Distribution of vetting assignments in the outcome and MRI safety tasks amongst consultant radiologists and within the consensus ground truth. There was substantial agreement between the two vetting radiologists in both the vetting outcome (Cohen's Kappa: 0.76) and MRI contra-indication detection (Cohen's Kappa: 0.68). Out of 120 synthetic referrals, the vetting radiologists agreed on the vetting outcome in 103 referrals and disagreed on 17 referrals. For MRI contra-indication detection, the vetting radiologists agreed on 115 referrals and disagreed on 5. Following review, we obtained a dataset with ground truth label distribution as shown in the final column of Table 1.

Model	Prompt type	Accuracy
OpenAI GPT-4	Instructions	0.54
	Instructions + guidance	0.82
	Instructions + guidance + examples	0.82
Anthropic Claude Opus	Instructions	0.68
	Instructions + guidance	0.82
	Instructions + guidance + examples	0.86
Google Gemini	Instructions	0.65
	Instructions + guidance	0.82
	Instructions + guidance + examples	0.78
All models (ensemble)	Instructions	0.65
	Instructions + guidance	0.88
	Instructions + guidance + examples	0.84

Table 2: Vetting outcome task results, showing accuracy of models when provided with varying levels of context in the prompt. In general, models which only received instruction tended to perform poorly compared to models which were given greater context in this task. Models which were given instruction and guidance performed similar to those given instructions, guidance and examples within the prompt. The best performing single best model was Anthropic Claude Opus when given instructions, guidance and examples which achieved an accuracy of 0.86.

Model	Prompt type	Precision	Recall	F1
OpenAI GPT-4	Instructions	1.00	0.78	0.88
	Instructions + guidance	0.62	0.56	0.59
	Instructions + guidance + examples	0.62	0.56	0.59
Anthropic Claude Opus	Instructions	0.42	0.89	0.57
	Instructions + guidance	0.50	0.56	0.53
	Instructions + guidance + examples	0.62	0.56	0.59
Google Gemini	Instructions	0.57	0.89	0.70
	Instructions + guidance	0.70	0.78	0.74
	Instructions + guidance + examples	0.47	0.78	0.58
All models (ensemble)	Instructions	0.80	0.89	0.84
	Instructions + guidance	0.62	0.56	0.59
	Instructions + guidance + examples	0.71	0.56	0.63

Table 3: MRI contra-indication identification task results, showing recall, precision and F1 scores of models when provided with varying levels of context in the prompt. In general models with instructions tended to outperform models with instructions and guidance or instructions with guidance and examples. No single model managed to achieve the highest score in both recall and precision however Open AI GPT-4 achieved the highest F1 score of 0.88. There appeared to be little difference between models which received instructions and guidance versus models which received instructions, guidance and examples.

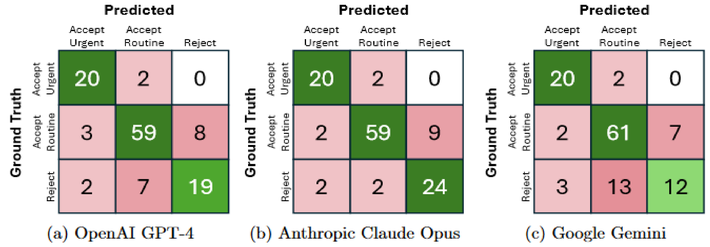


Figure 2: Confusion matrices of ground truth versus predicted assignments for the automatic referral vetting task for LLMs with full context. No model which received instructions, guidance and examples rejected a referral which had been vetted by the radiologists as urgent.

Conclusion

Summary statement: Large language models show promise in vetting MRI lumbar spine referrals, achieving up to 86% accuracy and identifying all urgent cases, potentially improving radiology workflow efficiency.

Key results:

- LLMs performed best when provided with local clinical guidelines, but adding a small number of example cases did not significantly improve performance.
- Models excelled at identifying MRI contraindications, with OpenAI GPT-4 achieving the highest F1 score of 0.88 using only basic instructions.
- LLMs vetted referrals significantly faster than radiologists (9.8 ± 1.0 minutes vs 135.0 ± 45.0 minutes for 120 referrals), suggesting potential for workload prioritisation (not shown on this poster).

Limitations

• Study design simplified vetting options to Accept (Routine/Urgent) and Reject, whereas real clinical practice includes additional categories like 'moderate' priority and options for follow-up imaging.

• Analysis was limited to text referrals only, not considering other patient data (like previous imaging) that radiologists might access when making vetting decisions.

• Use of synthetic data, while necessary, raises questions about practical deployment and data safeguarding when handling real patient information, particularly regarding third-party data sharing.

Data availability

All ground truth and prompts are available upon request.

References

[1] Royal College of Radiologists. RCR clinical radiology workforce census. 2022.

[2] Rebecca Smith-Bindman, Marilyn L Kwan, Emily C Marlow, Mary Kay Theis, Wesley Bolch, Stephanie Y Cheng, Erin JA Bowles, James R Duncan, Robert T Greenlee, Lawrence H Kushi, et al. Trends in use of medical imaging in us health care systems and in ontario, canada, 2000-2016. Jama, 322(9):843–856, 2019.

[3] Lisa C Adams, Daniel Truhn, Felix Busch, Avan Kader, Stefan M Niehues, Marcus R Makowski, and Keno K Bressem. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. Radiology, 307(4):e230725, 2023.

[4] A Infante, S Gaudino, F Orsini, A Del Ciello, C Gulli, B Merlino, L Natale, R Iezzi, and E Sala. Large language models (LLMs) in the evaluation of emergency radiology reports: performance of ChatGPT-4, perplexity, and bard. Clinical radiology, 79(2):102–106, 2024.