

Can LLMs Mimic Experienced Neurosurgeons in Real-World Spine Surgery Triage?

Mahvash Siavashpour[1,2(✉)], Erin McCabe[1], Andrew Nataraj[1], Niketh Pareek[3], Osmar Zaiane[1,2], and Douglas Gross[1]

[1] University of Alberta, Edmonton, Canada
siavashp@ualberta.ca
[2] Alberta Machine Intelligence Institute, Edmonton, Canada
[3] McGill University, Montreal, Canada

Abstract. Low back and neck pain are leading causes of disability, and some patients require surgical intervention to prevent long-term complications. Current triage workflows rely on MRI reports, questionnaires, and in-person assessment. However, many referred patients are ultimately managed non-surgically, contributing to long wait times and inefficient use of specialist resources. Automated decision support that distinguishes surgical from non-surgical cases could improve prioritization and care pathways.

This study investigates large language models (LLMs) for surgical triage using real-world data from the Spine Assessment Clinic in Edmonton, Canada. The dataset includes MRI reports and patient questionnaires, reflecting routine clinical inputs. We first evaluate conventional machine learning methods and find that these approaches perform poorly due to class imbalance, limited labelled data, and the complexity of radiology narratives. We then explore prompt-based use of open-source LLMs as an alternative to gradient-based training. LLMs outperform conventional models in this low-resource setting, particularly in unambiguous cases. Ablation analyses indicate that neither data modality alone is sufficient, but their combination yields the most reliable triage performance.

To the best of our knowledge, this work presents the first evaluation of LLMs for spine surgery triage on real-world clinical data. We benchmark multiple methods, examine multimodal integration strategies, and assess agreement with expert labels, highlighting both the promise and limitations of LLM-based clinical decision support.

Keywords: Large Language Models · Surgical Triage · Clinical Decision Support Systems

1 Introduction and Related Work

Low back and neck pain are common conditions, yet only a subset of patients require surgical intervention. In practice, many referred patients are managed

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
P. Andreev et al. (Eds.): AIME 2026, LNAI 16749, pp. 372–376, 2027.
https://doi.org/10.1007/978-3-032-30813-9_69

non-surgically, leading to unnecessary specialist consultations and prolonged wait times. At the Spine Assessment Clinic (SAC) in Edmonton, Canada, triage relies on diverse clinical inputs, including referral letters, MRI reports, patient questionnaires, and consultation findings. This motivates developing automated decision support systems to identify likely surgical candidates earlier. However, this task is challenging due to unstructured radiology narratives, high-dimensional and evolving questionnaire data, and limited, imbalanced labelled outcomes.

In recent years, machine learning has advanced clinical text processing [1]. However, triage using clinical text remains underexplored and often relies on curated datasets or narrow tasks. Prior work on the Spine Assessment Clinic triage showed that traditional NLP approaches can be outperformed by basic clinical variables, highlighting the difficulty of extracting surgical signals from radiology reports [3]. Multimodal approaches combining reports and questionnaires improve performance but still depend heavily on large labelled datasets [5]. LLMs can potentially offer an alternative by leveraging prior knowledge without task-specific training. Although recent studies examine LLMs for clinical reasoning and triage [2,4], they focus mainly on emergency or synthetic settings, leaving their effectiveness for real-world surgical decision support unclear.

In this work, we address surgical triage for spine care using a real clinical dataset with limited samples and heterogeneous inputs. We evaluate prompting-based use of open-source LLMs and compare them directly with conventional machine learning approaches. This setting allows us to study the feasibility of privacy-preserving, locally deployed LLM decision support for a nuanced clinical task that has received little prior attention.

2 Method

2.1 Data

We use de-identified MRI reports from 408 Spine Assessment Clinic patients, each paired with a 65-item self-assessment questionnaire. Of these, 69 (16.9%) were surgical and 339 (83.1%) non-surgical. The questionnaire captures demographics, clinical symptoms, and disability (Oswestry Disability Index), among others. The outcome is the neurosurgeon's final decision, made by one of five certified neurosurgeons. Decision codes were grouped into three categories: clear non-surgical (Codes 1–2), in-person assessments without surgery (Codes 3–4), and surgeon-involved decisions, including surgical (Codes 5–7) and non-surgical management (Codes 8–10). For consistency with clinical practices, decision codes 5, 6, and 7 are treated as surgical in the current work, while all the remaining codes are non-surgical.

2.2 LLM as Surgical Decision Support System

We propose a pipeline for leveraging the knowledge encoded in both the radiology reports and patient-completed questionnaires. We combine questionnaire and

report text with a prompt and use the LLM to classify patients as surgical or non-surgical. Here we describe the components of the pipeline:

To integrate questionnaire data into the model, each patient questionnaire is serialized into text following the `<field_name>: <field_value>` template. Questions with missing information are omitted from the serialized text. Radiology reports are processed by extracting the Findings and Impression sections [1], which are then combined with the serialized questionnaire to form different input representations for the LLM. To address the lack of fixed rules in spine surgery triage, this study evaluates multiple prompting strategies rather than relying on manually engineered prompts designed by medical experts, aiming to better align model predictions with expert decisions in a scalable way. For this purpose we evaluate three different prompting methods:

1. **Plain Prompting:** The model is provided with the patient's data and asked to directly classify the patient as surgical or non-surgical.

2. **Heuristic Prompting:** The prompt asks the LLM to act as a neurosurgeon and apply heuristics, which are manually derived from the literature, to determine whether surgery is needed. These heuristics include symptom and pain patterns, and imaging findings.

3. **Prompt Learning**: In this approach, the LLM is given a set of ground-truth examples and asked to identify patterns in the data and generate a prompt that includes information about how to do a surgical triage. We then use this prompt along with the patient's data to instruct a second LLM to make the final decision. Examples used in the prompt learning approach are selected randomly with an equal number of examples from each class.

Table 1. Qwen2.5-32B results compared to the baselines. Results are obtained under the best data combination for each model, averaged across folds (mean ± std).

Method	ROC AUC	Specificity	Sensitivity
Machine learning baseline			
Logistic Regression + BGE-M3	0.6172 ± 0.07	0.5958 ± 0.08	0.6385 ± 0.13
Fine-tuned RadBERT	0.5000 ± 0.00	1.0000 ± 0.00	0.0000 ± 0.00
Prompt-based LLM (Qwen2.5-32B)			
Plain Prompt	0.6681 ± 0.08	0.5219 ± 0.07	0.8143 ± 0.11
Heuristic Prompt	**0.6918 ± 0.09**	0.6133 ± 0.07	0.7703 ± 0.11
Prompt Learning	0.6172 ± 0.05	0.3629 ± 0.05	0.8714 ± 0.09

After constructing the prompt, the patient data and prompt are combined and fed into the LLM for classification. The model outputs a decision label along with a confidence score and key decision factors, improving interpretability and providing useful support for surgeons during triage.

3 Experimental Results

Table 1 reports the best-performing configuration for each model under its most informative data combination using 5-fold stratified cross-validation. The LLM temperature was set to 0 to eliminate stochasticity. Moreover, for brevity, we report only the strongest model per category, although more models were evaluated. The best results are obtained using the prompt-based method and when the model is instructed with explicit decision heuristics. Although prompt learning performs similarly to the baseline in ROC AUC, its high sensitivity suggests that letting the LLM craft decision rules from examples has resulted in a bias toward predicting the surgery class. In our experiments, limited training data and class imbalance prevented complex machine learning models, such as RadBERT [6], from learning robust patterns and resulted in them only predicting the majority class (negative), whereas simpler models could be trained more reliably. Additionally, Table 2 compares Qwen2.5-32B performance across three input modalities: MRI reports, questionnaires, and their combination. Questionnaire-only inputs show weak performance, suggesting limited predictive signal. Performance improves with radiology reports, but the best results come from combining both modalities, achieving the highest ROC AUC (0.69) and improved sensitivity. Overall, multimodal inputs are essential for accurate triage, and full reports outperform isolated sections.

Table 2. ROC AUC across input modalities. Abbreviations: Q = Questionnaire; RF = Report (findings); RI = Report (impression); R = Report (full); CF = Combined (findings); CI = Combined (impression); C = Combined (full).

	Q	RF	RI	R	CF	CI	C
ROC AUC	0.57	0.57	0.56	0.64	0.66	0.61	**0.69**

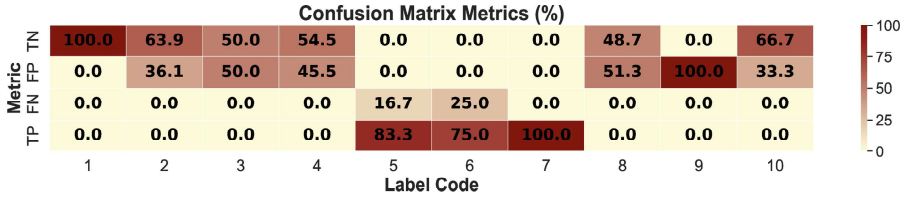


Fig. 1. Confusion matrix for Qwen2.5-32B (heuristic) on the full dataset. TN/FP/FN/TP denote true/false negatives/positives.

4 Discussion and Conclusions

After comparing the LLM results across patient categories in Fig. 1, we infer that performance is strongest for unambiguous surgical candidates (codes 5–7), indicating that the model reliably recognizes patients who ultimately require

imminent or urgent surgery. This supports the hypothesis that LLMs are effective at identifying clear surgical patterns. Performance is also strong for clear non-surgical cases decided without surgeon consultation (codes 1–2). However, the model performs poorly for borderline or uncertain clinical groups, which include patients seen by surgeons but managed conservatively or monitored over time. The results indicate that the model frequently predicts surgery when clinicians instead opt for observation or delayed surgical decision-making. This pattern implies that the model struggles with cases in which surgical candidacy depends on nuanced contextual factors, such as longitudinal response to treatment. The system therefore appears most suitable as a first-stage triage aid to help surgeons identify clearly surgical or clearly low-risk patients; it is intended only as a decision-support assistant, not a replacement for clinical judgment. Complex shared-decision cases still require expert review and may benefit from additional clinical context and more advanced prompt engineering.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Declaration of AI Assistance Technologies During Manuscript Preparation. The authors used ChatGPT (OpenAI) to paraphrase portions of the text. After using this technology, the authors have reviewed the results and take full responsibility for the content of the manuscript.

References

1. Han, Y., Yuan, Y., Mou, L.: A dual-view approach to classifying radiology reports by co-training. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 477–483. ELRA and ICCL, Torino, Italia, May 2024. https://aclanthology.org/2024.lrec-main.42/
2. Kim, S.H., et al.: Benchmarking the diagnostic performance of open-source LLMs in 1933 Eurorad case reports. npj Digit. Med. **8**(1), 97 (2025). https://doi.org/10.1038/s41746-025-01488-3
3. Krebs, B., et al.: Developing a triage predictive model for access to a spinal surgeon using clinical variables and natural language processing of radiology reports. Eur. Spine J. **35**(3), 1330–1343 (2026). https://doi.org/10.1007/s00586-023-07552-4
4. Lee, J., et al.: From promising capability to pervasive bias: assessing large language models for emergency department triage. arXiv preprint arXiv:2504.16273 (2025)
5. Siavashpour, M., McCabe, E., Nataraj, A., Pareek, N., Zaiane, O., Gross, D.: A multimodal deep learning ensemble framework for building a spine surgery triage system. Stud. Health Technol. Inf. **329**, 1824–1825 (2025). https://doi.org/10.3233/SHTI251233
6. Yan, A., et al.: RadBERT: adapting transformer-based language models to radiology. Radiol. Artif. Intell. **4**(4), e210258 (2022). https://doi.org/10.1148/ryai.210258