Original research

Evaluating local open-source large language models for data extraction from unstructured reports on mechanical thrombectomy in patients with ischemic stroke

Aymen Meddeb ,[1,2] Philipe Ebert,[1] Keno Kyrill Bressem,[3] Dmitriy Desser,[1] Andrea Dell'Orco,[1] Georg Bohner,[1] Justus F Kleine,[1] Eberhard Siebert ,[1] Nils Grauhan,[4] Marc A Brockmann,[4] Ahmed Othman,[4] Michael Scheel,[1] Jawed Nawabi[1]

[1]Department of Neuroradiology, Charité Universitätsmedizin Berlin, Berlin, Germany
[2]Department of Neuroradiology, CHU Reims Imagerie Médicale, Reims, Champagne-Ardenne, France
[3]German Heart Center Munich, Technical University of Munich, Munchen, Germany
[4]Department of Neuroradiology, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany

Correspondence to
Dr Aymen Meddeb; aymen.meddeb@charite.de

Received 5 June 2024
Accepted 9 July 2024

To cite: Meddeb A, Ebert P, Bressem KK, *et al*. *J NeuroIntervent Surg* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/jnis-2024-022078

ABSTRACT

Background A study was undertaken to assess the effectiveness of open-source large language models (LLMs) in extracting clinical data from unstructured mechanical thrombectomy reports in patients with ischemic stroke caused by a vessel occlusion.

Methods We deployed local open-source LLMs to extract data points from free-text procedural reports in patients who underwent mechanical thrombectomy between September 2020 and June 2023 in our institution. The external dataset was obtained from a second university hospital and comprised consecutive cases treated between September 2023 and March 2024. Ground truth labeling was facilitated by a human-in-the-loop (HITL) approach, with time metrics recorded for both automated and manual data extractions. We tested three models—Mixtral, Qwen, and BioMistral—assessing their performance on precision, recall, and F1 score across 15 clinical categories such as National Institute of Health Stroke Scale (NIHSS) scores, occluded vessels, and medication details.

Results The study included 1000 consecutive reports from our primary institution and 50 reports from a secondary institution. Mixtral showed the highest precision, achieving 0.99 for first series time extraction and 0.69 for occluded vessel identification within the internal dataset. In the external dataset, precision ranged from 1.00 for NIHSS scores to 0.70 for occluded vessels. Qwen showed moderate precision with a high of 0.85 for NIHSS scores and a low of 0.28 for occluded vessels. BioMistral had the broadest range of precision, from 0.81 for first series times to 0.14 for medication details. The HITL approach yielded an average time savings of 65.6% per case, with variations from 45.95% to 79.56%.

Conclusion This study highlights the potential of using LLMs for automated clinical data extraction from medical reports. Incorporating HITL annotations enhances precision and also ensures the reliability of the extracted data. This methodology presents a scalable privacy-preserving option that can significantly support clinical documentation and research endeavors.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Large language models (LLMs) have shown promise in various natural language processing tasks including extracting data from unstructured texts.

⇒ Clinical data extraction from medical reports is a critical task that can benefit from automation due to the labor-intensive and time-consuming nature of manual extraction.

WHAT THIS STUDY ADDS

⇒ This study shows the specific application of open-source LLMs in extracting clinical data from unstructured mechanical thrombectomy reports.

⇒ The integration of a human-in-the-loop (HITL) approach significantly enhances the precision and reliability of the extracted data.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Open-source LLMs, when combined with HITL annotations, offer a scalable and privacy-preserving solution for clinical data extraction, enhancing the efficiency and accuracy of clinical documentation and research.

INTRODUCTION

Large language models (LLMs) are artificial intelligence (AI) systems that understand and generate human-like natural language responses to text prompts.[1–3] These models, trained on vast datasets, have shown remarkable clinical reasoning capabilities[4–6] in passing medical licensing examinations[7 8] and generating prevention and treatment recommendations for various conditions including cardiovascular disease[9 10] and breast cancer.[11] They can produce clinical notes,[3] generate radiology reports,[12 13] and even assist in writing research articles.[14–16]

Although GPT-4 has been effectively used for text mining from unstructured medical data in radiology[17] and neurology,[18 19] the application of commercial LLMs in medicine raises significant privacy issues, with the important need to uphold the strict standards of data privacy and security inherent to the clinical context. In addition, reliability and usability are important issues that must be addressed when using LLM. Users, especially in the medical field, must be able to rely on the accuracy and consistency of the model, which requires ongoing refining, testing, and evaluation to ensure that the system is delivering accurate outputs. However, commercial models behind API can be updated by the provider and drastically change their behavior, which can pose a substantial risk for clinical workflows.

In neuroradiology, the production of accurate and detailed notes after interventional procedures are crucial.[20] This documentation must meticulously describe the intervention, reflecting the highly individualized nature of each procedure. It should detail the indication for the procedure, enumerate the technical steps undertaken, list the materials and medications used, address potential complications, and report the outcome of the intervention.[21] This level of documentation is essential for upholding high care standards and also for supporting efficient patient discharge, enabling quality assessments, and enhancing clinical research.

However, this highly individualized nature of procedural notes hampers the use of generated data. The diversity in documentation practices and writing styles poses a significant challenge for structuring the data for research purposes or integrating it into national registries. Moreover, the variability in detail and terminology used complicates the task of standardizing data for comparative studies or broader analysis.

This is where open-source LLMs present a compelling solution.[22–24] By operating fully locally, these models ensure that all data processing is confined to the hospital's internal devices and designated servers without the need for external internet connectivity. This mode of operation mitigates the risk of data breaches and aligns with the principles of patient data privacy.

This work aims to explore the potential of open-source local LLMs in extracting accurate information from procedural reports of mechanical thrombectomy in patients with ischemic stroke and accelerating annotation for medical information extraction to fully leverage the rich data contained in procedural notes for quality assurance, research, and regulatory purposes.

METHODS

Patient population

Our internal dataset encompasses consecutive reports from patients who underwent mechanical thrombectomy for acute ischemic stroke between September 2020 and June 2023 in a university hospital with a comprehensive stroke center. The external dataset encompasses consecutive reports from patients treated in a second university hospital between September 2023 and March 2024. All collected data adhered to the principles outlined in the Declaration of Helsinki. Patient data were anonymized to ensure privacy and confidentiality, in line with the stringent data protection requirements of clinical research.

Extraction pipeline and prompt structure

To adapt a generalized LLM for our specific task of information extraction, we employed an in-context learning strategy.[25] This method involves crafting precise prompts to provide the model with clear instructions and context, enhancing its ability to perform complex tasks. We developed an automated pipeline to process the reports, beginning with the creation of a JavaScript Object Notation (JSON) template. This template defined the 15 data points we aimed to extract from the thrombectomy reports: National Institutes of Health Stroke Scale (NIHSS) score, symptom onset, occluded vessel, occlusion side, used materials, medication, complications, outcome, Thrombolysis in Cerebral Infarction (TICI) score, area dose product, fluoroscopy time, arrival time, puncture time, first series time, and artery opening time. For each data point a detailed list of instructions was provided to clarify definitions and specifics. Finally, a prompt was crafted to analyze the reports, extract the necessary data, and populate the JSON template. The system was also configured to process texts primarily in German. All notebooks and prompts are publicly available in GitHub (https://github.com/Meddebma/AI_4_Medicine/blob/main/Thrombectomy_LLM_Extraction.ipynb).

Table 1 Overview of the used open-source models

Model	Qwen	Mixtral	BioMistral	Phi-2
Parameters	72 b	8×7 b	7 b	2 b
Architecture	Decoder-only Transformer	Decoder-only Transformer	Decoder-only Transformer	Encoder-decoder Transformer
Manufacturer	Alibaba Cloud	Mistral AI	Mistral AI	Microsoft

b, billion.

Models

For information extraction we implemented four state-of-the-art open-source LLMs, each varying in the number of parameters: Qwen with 72 billion, Mixtral configured as 8 clusters of 7 billion each, and BioMistral with 7 billion parameters. Each model was deployed locally to align with stringent data security measures and our commitment to patient privacy. Additionally, we tested the Phi-2 model with 2 billion parameters; however, it did not produce reasonable outputs and was subsequently excluded from further analysis. Table 1 provides an overview of the used models.

Human-in-the-loop (HITL) annotation

Due to the extensive volume of cases, we adopted a human-in-the-loop (HITL) annotation strategy to establish ground truth for our dataset. Rather than creating annotations from scratch, clinical experts refined the outputs from our strongest model, Mixtral. This refinement process required a detailed review and adjustment by four clinical experts comprising one student, one radiology resident, and two board-certified radiologists specializing in neurointerventions to ensure the outputs were accurate and clinically relevant. The combined effort amounted to approximately 80 hours of work. The resulting annotated dataset then served as a benchmark to evaluate the performance of subsequent models. This workflow is shown in figure 1.

To determine the most effective model for our HITL annotation process we initially conducted a series of preliminary evaluations. These evaluations assessed the precision of each model on consistently reported data elements within a small subset of our data (n=20). Based on these initial results, Mixtral emerged as the superior model, particularly excelling in the extraction of critical data points such as used materials data. This decision to use Mixtral as the basis for our HITL approach allowed us to begin the augmentation process from the most reliable automated baseline, ensuring the high quality of our ground truth data.

To assess the time reduction afforded by using an LLM in data extraction tasks we meticulously recorded the time required for the manual extraction of 30 reports. Additionally, we measured the time taken for the LLM to extract data and the subsequent

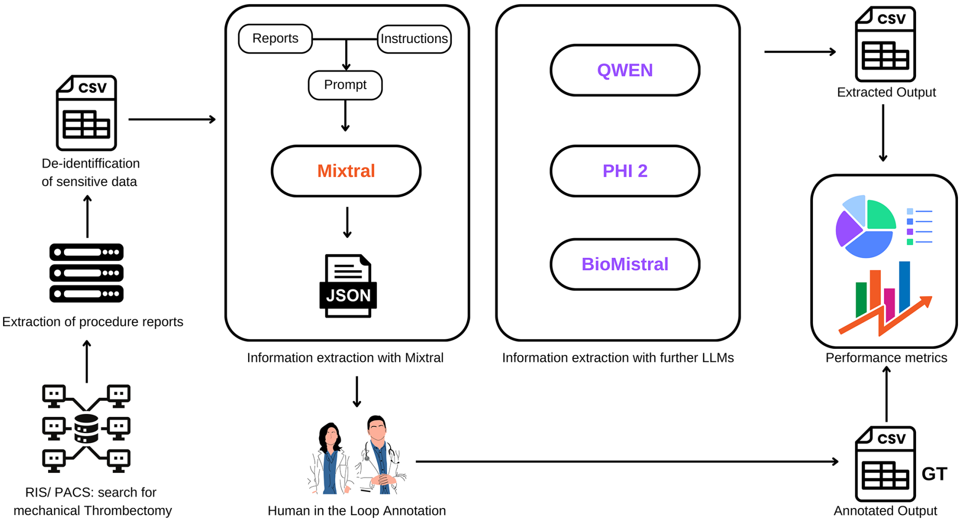
CSV
De-identiffication of sensitive data
Extraction of procedure reports
RIS/ PACS: search for mechanical Thrombectomy
Reports
Instructions
Prompt
Mixtral
JSON
Information extraction with Mixtral
Human in the Loop Annotation
QWEN
PHI 2
BioMistral
Information extraction with further LLMs
CSV
Extracted Output
Performance metrics
CSV
GT
Annotated Output

Figure 1 Workflow design for information extraction.

time needed for manual corrections of the LLM-extracted information. To statistically evaluate the significance of the observed differences in time between manual and assisted methods, we conducted a paired t-test. The α level for determining statistical significance was set at p=0.05. Furthermore, we quantified the efficiency gains from LLM assistance by calculating the percentage of time saved.

Handling missing data and extraction failures

The used LLMs were prompted to explicitly identify and label missing data points in the thrombectomy reports as 'not available' or 'not applicable'. This approach was employed to ensure clarity and prevent the generation of incorrect or fabricated data. Furthermore, we incorporated a feedback loop to reinitiate the extraction process after initial failures. This system was crucial for identifying reports where data extraction was not feasible, such as those involving venous sinus thrombectomy or spontaneous recanalization, ensuring accurate data handling.

Evaluation metrics

To comprehensively evaluate the performance of the LLMs we used a range of metrics designed to assess various aspects of the output quality of the models—namely, precision, recall, and F1 score. We performed all statistical analyses using the Pandas and SciPy libraries in Python (Version 3.12.1) and plots were created using RStudio (R Version 4.3.2).

As the column of used materials includes a list of materials separated with a comma in the output JSON file, we opted for token-based metrics to count the number of extracted items as it is more adapted than determining the accuracy of the whole list.

RESULTS

Study population

Initially, 1026 reports were retrieved using our radiology information system. Eighteen reports were excluded due to spontaneous recanalization or absence of intracranial occlusion and eight further reports were excluded due to venous sinus thrombectomy. Included reports were written by seven different neurointerventionalists. The external dataset comprised 50 reports on mechanical thrombectomy performed at a second university hospital.

Extracted information

Our evaluation of LLMs for extracting information from unstructured thrombectomy reports showed variable effectiveness across diverse metrics. Among the models tested, Mixtral achieved the highest performance with precision values ranging from 0.99 for first series time to 0.69 for occluded vessel data. The Qwen model showed moderate performance with precision scores from 0.85 for the NIHSS score to 0.28 for occluded vessels. Despite its specialization in medical tasks, the BioMistral model had the lowest precision, with scores peaking at 0.81 for first series time and dipping to 0.14 for medication data.

Notably, all models performed well in extracting explicit data facilitated by the use of an integrated template within the reports. For instance, precision for NIHSS score was high across models (Mixtral: 0.98, Qwen: 0.85, BioMistral: 0.79). Similarly, scores for puncture time (Mixtral: 0.98, Qwen: 0.82, BioMistral: 0.82), first series time (Mixtral: 0.99, Qwen: 0.81, BioMistral: 0.81), and artery opening time (Mixtral: 0.98, Qwen: 0.79, BioMistral: 0.78) indicated strong performance. However, precision was notably lower for occluded vessel extraction in the Mixtral model (0.69) and for medication details in both the BioMistral (0.14) and Qwen models (0.28).

For the external dataset we used only the Mixtral model. This model showed high precision across various data points ranging from a perfect 1.00 for the NIHSS score to 0.70 for occluded vessels.

The detailed performance metrics for the internal and external datasets are shown in table 2 and table 3, respectively. Figure 2 visually illustrates the precision values for each model, including error bars that highlight the variability in performance across the tested data points.

Human-in-the-loop annotation

The analysis showed that the mean time required for manual data extraction was 186.95 s (range 37–401 s), the mean time for initial data extraction by the LLM was 4.33 s (range 2–6 s), and the mean time needed for manual corrections of the LLM-extracted information was 59.63 s (range 20–103 s). These efficiencies resulted in an average time savings of 65.6% per case (range 45.95–79.56%). The time difference was statistically significant (p<0.05).

Table 2 Performance metrics for the internal dataset

Model	Mixtral			BioMistral			Qwen		
Metrics	**Precision**	**Recall**	**F1 score**	**Precision**	**Recall**	**F1 score**	**Precision**	**Recall**	**F1 score**
NIHSS score	0.98 (0.97 to 0.99)	0.97 (0.96 to 0.98)	0.97 (0.96 to 0.98)	0.79 (0.70 to 0.88)	0.62 (0.55 to 0.68)	0.63 (0.56 to 0.70)	0.85 (0.82 to 0.88)	0.85 (0.83 to 0.88)	0.84 (0.82 to 0.87)
Symptom onset	0.78 (0.75 to 0.81)	0.83 (0.81 to 0.86)	0.80 (0.77 to 0.83)	0.48 (0.40 to 0.56)	0.47 (0.41 to 0.54)	0.46 (0.39 to 0.54)	0.58 (0.54 to 0.62)	0.42 (0.39 to 0.46)	0.47 (0.44 to 0.50)
Occluded vessel	0.69 (0.66 to 0.72)	0.72 (0.69 to 0.75)	0.69 (0.66 to 0.71)	0.36 (0.28 to 0.45)	0.36 (0.29 to 0.43)	0.34 (0.27 to 0.42)	0.51 (0.46 to 0.56)	0.32 (0.29 to 0.34)	0.35 (0.32 to 0.39)
Occlusion side	0.87 (0.84 to 0.90)	0.86 (0.84 to 0.88)	0.84 (0.82 to 0.87)	0.65 (0.51 to 0.75)	0.67 (0.60 to 0.73)	0.61 (0.53 to 0.69)	0.71 (0.67 to 0.74)	0.74 (0.71 to 0.77)	0.72 (0.69 to 0.75)
Used materials	0.94 (0.92 to 0.95)	0.92 (0.91 to 0.93)	0.93 (0.91 to 0.94)	0.56 (0.52 to 0.59)	0.56 (0.53 to 0.59)	0.55 (0.52 to 0.59)	0.57 (0.55 to 0.59)	0.55 (0.53 to 0.57)	0.56 (0.54 to 0.58)
Medication	0.88 (0.86 to 0.90)	0.84 (0.81 to 0.86)	0.84 (0.81 to 0.86)	0.14 (0.08 to 0.19)	0.22 (0.17 to 0.28)	0.14 (0.10 to 0.20)	0.28 (0.23 to 0.33)	0.20 (0.18 to 0.23)	0.21 (0.19 to 0.24)
Complications	0.87 (0.85 to 0.90)	0.90 (0.88 to 0.91)	0.88 (0.86 to 0.90)	0.45 (0.30 to 0.56)	0.17 (0.12 to 0.23)	0.13 (0.08 to 0.19)	0.40 (0.35 to 0.44)	0.20 (0.18 to 0.23)	0.26 (0.23 to 0.29)
Outcome	0.80 (0.77 to 0.83)	0.81 (0.78 to 0.83)	0.79 (0.76 to 0.81)	0.31 (0.15 to 0.48)	0.08 (0.04 to 0.12)	0.10 (0.05 to 0.14)	0.51 (0.42 to 0.58)	0.07 (0.06 to 0.09)	0.10 (0.07 to 0.12)
TICI score	0.91 (0.89 to 0.93)	0.89 (0.87 to 0.91)	0.89 (0.87 to 0.91)	0.54 (0.46 to 0.63)	0.65 (0.58 to 0.72)	0.59 (0.51 to 0.66)	0.50 (0.46 to 0.54)	0.62 (0.59 to 0.65)	0.54 (0.50 to 0.57)
Area dose product	0.98 (0.96 to 0.99)	0.98 (0.97 to 0.99)	0.98 (0.97 to 0.99)	0.48 (0.41 to 0.56)	0.53 (0.46 to 0.59)	0.50 (0.43 to 0.57)	0.75 (0.72 to 0.78)	0.79 (0.77 to 0.81)	0.77 (0.73 to 0.80)
Fluoroscopy time	0.97 (0.96 to 0.98)	0.97 (0.96 to 0.98)	0.97 (0.96 to 0.98)	0.40 (0.33 to 0.47)	0.40 (0.33 to 0.47)	0.39 (0.33 to 0.46)	0.57 (0.54 to 0.61)	0.62 (0.59 to 0.65)	0.58 (0.55 to 0.62)
Arrival time	0.96 (0.94 to 0.97)	0.96 (0.94 to 0.97)	0.96 (0.94 to 0.97)	0.72 (0.64 to 0.80)	0.65 (0.58 to 0.71)	0.66 (0.60 to 0.73)	0.69 (0.66 to 0.72)	0.66 (0.63 to 0.68)	0.64 (0.61 to 0.67)
Puncture time	0.98 (0.97 to 0.99)	0.99 (0.98 to 0.99)	0.98 (0.98 to 0.99)	0.80 (0.73 to 0.85)	0.76 (0.71 to 0.82)	0.78 (0.72 to 0.83)	0.82 (0.79 to 0.85)	0.78 (0.76 to 0.81)	0.79 (0.76 to 0.81)
First series time	0.99 (0.98 to 1.00)	0.99 (0.98 to 1.00)	0.99 (0.98 to 0.99)	0.81 (0.75 to 0.87)	0.79 (0.72 to 0.84)	0.80 (0.73 to 0.85)	0.81 (0.79 to 0.84)	0.81 (0.78 to 0.83)	0.80 (0.77 to 0.82)
Artery opening time	0.98 (0.97 to 0.99)	0.98 (0.97 to 0.99)	0.98 (0.97 to 0.99)	0.78 (0.72 to 0.85)	0.77 (0.70 to 0.82)	0.77 (0.71 to 0.83)	0.79 (0.76 to 0.82)	0.77 (0.74 to 0.79)	0.76 (0.73 to 0.79)

Data are shown as mean (95% Confidence Interval, CI).
NIHSS, National Institutes of Health Stroke Scale; TICI, Thrombolysis in Cerebral Infarction.

Table 3 Performance metrics using Mixtral model for the external dataset

Metric	Precision	Recall	F1 score
NIHSS score	1.000 (1.000 to 1.000)	1.000 (1.000 to 1.000)	1.000 (1.000 to 1.000)
Symptom onset	1.000 (1.000 to 1.000)	1.000 (1.000 to 1.000)	1.000 (1.000 to 1.000)
Occluded vessel	0.705 (0.413 to 0.910)	0.643 (0.479 to 0.840)	0.618 (0.402 to 0.810)
Occlusion side	0.968 (0.908 to 1.000)	0.918 (0.800 to 1.000)	0.936 (0.834 to 1.000)
Used materials	1.000 (1.000 to 1.000)	0.988 (0.964 to 1.000)	0.993 (0.979 to 1.000)
Medication	0.961 (0.880 to 1.000)	0.919 (0.800 to 1.000)	0.938 (0.828 to 1.000)
Complications	0.960 (0.880 to 1.000)	0.962 (0.880 to 1.000)	0.961 (0.880 to 1.000)
Outcome	0.877 (0.696 to 1.000)	0.879 (0.720 to 1.000)	0.875 (0.731 to 1.000)
TICI score	0.983 (0.880 to 1.000)	0.960 (0.880 to 1.000)	0.967 (0.880 to 1.000)

Data are shown as mean (95% Confidence Interval, CI).
NIHSS, National Institutes of Health Stroke Scale; TICI, Thrombolysis in Cerebral Infarction.

DISCUSSION

Extracting meaningful data from unstructured medical text is both challenging and essential for data analysis and research. Our study highlights the feasibility of automated data extraction from thrombectomy reports in patients with stroke using open-source LLMs within a secure local environment that respects patient data privacy. The Mixtral 8×7b instruct model demonstrated high performance, with precision values ranging from 0.99 for first series time to 0.68 for occluded vessel. Our results indicate that LLMs can effectively contribute to medical research by streamlining data processing while safeguarding sensitive information.

Recent studies underscore the high efficacy of LLMs in extracting both implicit and explicit data from unstructured text. Dagdelen *et al* explored the use of LLMs for material science data extraction using models with a JSON output schema. Their findings show a clear superiority of LLMs over traditional natural language processing methods, highlighting significant time savings achieved through the HITL annotation technique.[26] Similarly, a study by Goel *et al* showed that LLMs could significantly accelerate medical data extraction. Their approach, which also used HITL annotations, reduced time costs by an average of 42% compared with traditional annotation methods from scratch.[27] In our study, HITL reduced time by an average of 65%.

For mechanical thrombectomy procedures in patients with stroke, various studies have demonstrated the effectiveness of extracting procedural details from free-text reports. Yu *et al* used a traditional natural language processing approach to detect large vessel occlusion in radiologic reports, achieving an accuracy of 97.3%.[28] Gunter *et al* also reported accuracy greater than 90% in identifying different stroke characteristics from radiology reports.[29] More recently, Lehmen *et al* used GPT-3.5 and GPT-4 to extract data from 100 mechanical thrombectomy reports, with a correctness rate of 94% across data points and performance varying between 61% and 100% per category.[30] However, a significant limitation of this approach is the potential compromise of data privacy. In contrast, our automated local LLM pipeline achieved comparable results in 1000 reports while ensuring a completely secure environment for data processing, thus respecting patient data privacy. This approach maintains high performance in data extraction and also upholds strict data protection standards.

The accuracy and precision of the different models revealed varying degrees across extracted data points. For instance, the TICI score and NIHSS score showed high precision across all models, indicating robust model performance to extract explicit data. In contrast, 'used materials', 'medication', and 'complications' showed lower precision, suggesting higher complexity of these implicit data. For example, the Mixtral model correctly identified a peri-interventional distal embolus with peripheric small vessel occlusion or a failure of closure device with a subsequent groin hematoma as complications whereas these were not detected by the BioMistral and Qwen models.

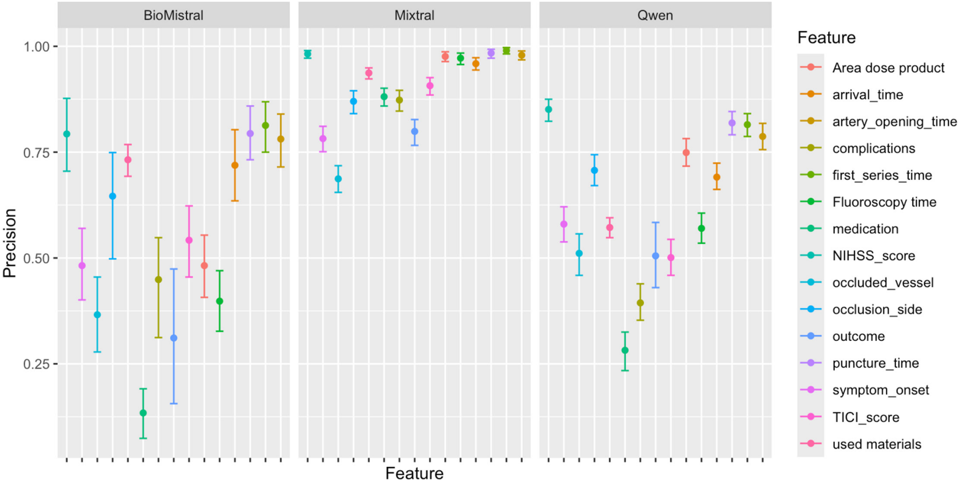

Figure 2 Precision for prediction of data points with different parameter size models.

We observed moderate performance in extracting data for the 'occluded vessel' category, attributable to two main factors. First, the lack of standardized nomenclature for vessels poses a significant challenge; terms like 'distal ICA', 'carotid terminus', and 'supraophthalmic segment' refer to the same location but are labeled differently by neurointerventionalists. Second, there are often discrepancies between the occluded vessel as noted in the clinical history and the requested procedure section, which typically describe the occlusion identified on CT or MRI scans, and the results section which reflects the occlusion detected during angiography. These inconsistencies contribute to the difficulties in accurately extracting and interpreting data regarding occluded vessels from procedural reports.

Our study has some limitations. First, its retrospective nature may limit the generalizability of the results. Second, the results are based on the performance of the models in extracting data from German reports and therefore may not be directly applicable to reports in other languages. However, especially for English text, one can assume that the model performance will even increase as the training data of the models consists mainly of English texts. Last, variability in the quality and consistency of the input data, such as differences in terminology, formatting, or detail level in the reports, can affect the performance of the models; however, our model showed a stable performance in extracting data from the external dataset.

CONCLUSION

Our findings show that an automated pipeline for data extraction from procedural reports using local open-source LLMs is both feasible and effective, achieving high performance levels. Furthermore, integrating an HITL annotation process can significantly reduce the time cost while ensuring reliable results.

Acknowledgements AM is fellow of the BIH Charité Digital Clinician Scientist Program funded by the Charité – Universitätsmedizin Berlin and the Berlin Institute of Health at Charité (BIH).

Contributors AM and JN designed the study. PE, KKB, DD, AD, GB, JFK, ES and MS gathered and critically reviewed the data. NG, MAB and AO gathered and critically reviewed the external dataset, AM performed statistical analyses and drafted the manuscript. PE, KKB, DD, AD, GB, JFK, ES, MS, NG, MAB, AO and JN critically reviewed the manuscript. All authors approved the final version of the manuscript. AM is responsible for the overall content as guarantor.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This retrospective study was approved by the ethics committee of the Charité University Hospital in Berlin (No. EA4/062/20).The requirement for informed consent was waived due to the retrospective design of the study.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. All notebooks and prompts used in this study are publicly available at GitHub (https://github.com/Meddebma/AI_4_Medicine/blob/main/Thrombectomy_LLM_Extraction.ipynb).

ORCID iDs
Aymen Meddeb http://orcid.org/0000-0001-6537-9419
Eberhard Siebert http://orcid.org/0000-0001-7395-6546

REFERENCES

1 Liu Y, Han T, Ma S, *et al*. Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiol* 2023;1:100017.
2 Li J, Dada A, Kleesiek J, *et al*. ChatGPT in healthcare: a taxonomy and systematic review. *Health Informatics* 2023.03.30.23287899 [Preprint] 2023.
3 Cascella M, Montomoli J, Bellini V, *et al*. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst* 2023;47:33.
4 Lee P, Bubeck S, Petro JB. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. *N Engl J Med* 2023;388:1233–9.
5 Truhn D, Reis-Filho JS, Kather JN. Large language models should be used as scientific reasoning engines, not knowledge databases. *Nat Med* 2023;29:2983–4.
6 Xie S, Zhao W, Deng G, *et al*. Utilizing ChatGPT as a scientific reasoning engine to differentiate conflicting evidence and summarize challenges in controversial clinical questions. *J Am Med Inform Assoc* 2024;31:1551–60.
7 Kung TH, Cheatham M, Medenilla A, *et al*. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2:e0000198.
8 Han T, Adams LC, Bressem KK, *et al*. Comparative analysis of multimodal large language model performance on clinical vignette questions. *JAMA* 2024;331:1320–1.
9 Sarraju A, Bruemmer D, Van Iterson E, *et al*. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA* 2023;329:842–4.
10 Zhu L, Mou W, Wu K, *et al*. Multimodal ChatGPT-4V for ECG interpretation: promise and limitations. *J Med Internet Res* 2023;26:e54607.
11 Haver HL, Ambinder EB, Bahl M, *et al*. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. *Radiology* 2023;307:e230424.
12 Elkassem AA, Smith AD. Potential use cases for ChatGPT in radiology reporting. *Am J Roentgenol* 2023;221:373–6.
13 Amin KS, Davis MA, Doshi R, *et al*. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for simplifying radiology reports. *Radiology* 2023;309:e232561.
14 Hosseini M, Rasmussen LM, Resnik DB. Using AI to write scholarly publications. *Account Res* 2023;1–9.
15 Kadi G, Aslaner MA. Exploring ChatGPT's abilities in medical article writing and peer review. *Croat Med J* 2024;65:93–100.
16 Safrai M, Orwig KE. Utilizing artificial intelligence in academic writing: an in-depth evaluation of a scientific review on fertility preservation written by ChatGPT-4. *J Assist Reprod Genet* 2024;1–10.
17 Adams LC, Truhn D, Busch F, *et al*. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. *Radiology* 2023;307:e230725.
18 Zhang H, Jethani N, Jones S, *et al*. Evaluating large language models in extracting cognitive exam dates and scores. *Health Informatics* [Preprint].
19 Du X, Novoa-Laurentiev J, Plasaek JM, *et al*. Enhancing early detection of cognitive decline in the elderly: a comparative study utilizing large language models in clinical notes. *Health Informatics* [Preprint].
20 Smouse HB, Harty P. Paperwork for the busy interventionalist: the basic six. *Semin Intervent Radiol* 2006;23:319–28.
21 Mahnken AH, Boullosa Seoane E, Cannavale A, *et al*. CIRSE clinical practice manual. *Cardiovasc Intervent Radiol* 2021;44:1323–53.
22 Wu C, Lin W, Zhang X, *et al*. PMC-LLaMA: toward building open-source language models for medicine. *J Am Med Inform Assoc* 2024.:ocae045.
23 Bai J, Kamatchinathan S, Kundu DJ, *et al*. Open-source large language models in action: a bioinformatics chatbot for PRIDE database. *Proteomics* 2024.:e2400005.
24 Le Guellec B, Lefèvre A, Geay C, *et al*. Performance of an open-source large language model in extracting information from free-text radiology reports. *Radiol Artif Intell* 2024;6:e230364.
25 Brown TB, Mann B, Ryder N, *et al*. Language models are few-shot learners. *arXiv* 2020.
26 Dagdelen J, Dunn A, Lee S, *et al*. Structured information extraction from scientific text with large language models. *Nat Commun* 2024;15:1418.
27 Goel A, Gueta A, Gilon O, *et al*. LLMs accelerate annotation for medical information extraction. *arXiv* 2023.
28 Yu AYX, Liu ZA, Pou-Prom C, *et al*. Automating stroke data extraction from free-text radiology reports using natural language processing: instrument validation study. *JMIR Med Inform* 2021;9:e24381.
29 Gunter D, Puac-Polanco P, Miguel O, *et al*. Rule-based natural language processing for automation of stroke data extraction: a validation study. *Neuroradiology* 2022;64:2357–62.
30 Lehnen NC, Dorn F, Wiest IC, *et al*. Data extraction from free-text reports on mechanical thrombectomy in acute ischemic stroke using ChatGPT: a retrospective analysis. *Radiology* 2024;311:e232741.