UC Davis
UC Davis Previously Published Works

Title

Performance of an Open-Source Large Language Model in Extracting Information from Free-Text Radiology Reports.

Permalink

https://escholarship.org/uc/item/6nb547t0

Journal

Radiology: Artificial Intelligence, 6(4)

Authors

Le Guellec, Bastien

Lefèvre, Alexandre

Geay, Charlotte

et al.

Publication Date

2024-07-01

DOI

10.1148/ryai.230364

Peer reviewed

Performance of an Open-Source Large Language Model in Extracting Information from Free-Text Radiology Reports

Bastien Le Guellec, MS • Alexandre Lefèvre, MS • Charlotte Geay, MSc • Lucas Shorten, MSc • Cyril Bruge, MD • Lotfi Hacein-Bey, MD • Philippe Amouyel, MD, PhD • Jean-Pierre Pruvo, MD, PhD • Gregory Kuchcinski, MD, PhD • Aghiles Hamroun, MD, PhD

From the Department of Neuroradiology (B.L.G., A.L., C.B., J.P.P., G.K.), Department of Public Health (B.L.G., P.A., A.H.), and INclude Health Data Warehouse (C.G., L.S.), CHU Lille–Université Lille, Rue Emile Laine, 59000 Lille, France; Department of Radiology, UC Davis Health, Sacramento, Calif (L.H.B.); Université Lille, INSERM, CHU Lille, Institut Pasteur de Lille, U1167-RID-AGE - Facteurs de risque et déterminants moléculaires des maladies liées au vieillissement, Lille, France (P.A., A.H.); INSERM, U1172–LilNCog-Lille Neuroscience & Cognition, Université Lille, Lille, France (J.P.P., G.K.); and UAR 2014-US 41-PLBS–Plateformes Lilloises en Biologie & Santé, Université Lille, Lille, France (J.P.P., G.K.). Received August 31, 2023; revision requested October 12; final revision received March 8, 2024; accepted April 23. **Address correspondence to** B.L.G. (email: *bastien.leguellec.etu@univ-lille.fr*).

Purpose: To assess the performance of a local open-source large language model (LLM) in various information extraction tasks from real-life emergency brain MRI reports.

Materials and Methods: All consecutive emergency brain MRI reports written in 2022 from a French quaternary center were retrospectively reviewed. Two radiologists identified MRI scans that were performed in the emergency department for headaches. Four radiologists scored the reports' conclusions as either normal or abnormal. Abnormalities were labeled as either headache-causing or incidental. Vicuna (LMSYS Org), an open-source LLM, performed the same tasks. Vicuna's performance metrics were evaluated using the radiologists' consensus as the reference standard.

Results: Among the 2398 reports during the study period, radiologists identified 595 that included headaches in the indication (median age of patients, 35 years [IQR, 26–51 years]; 68% [403 of 595] women). A positive finding was reported in 227 of 595 (38%) cases, 136 of which could explain the headache. The LLM had a sensitivity of 98.0% (95% CI: 96.5, 99.0) and specificity of 99.3% (95% CI: 98.8, 99.7) for detecting the presence of headache in the clinical context, a sensitivity of 99.4% (95% CI: 98.3, 99.9) and specificity of 98.6% (95% CI: 92.2, 100.0) for the use of contrast medium injection, a sensitivity of 96.0% (95% CI: 92.5, 98.2) and specificity of 98.9% (95% CI: 97.2, 99.7) for study categorization as either normal or abnormal, and a sensitivity of 88.2% (95% CI: 81.6, 93.1) and specificity of 73% (95% CI: 62, 81) for causal inference between MRI findings and headache.

Conclusion: An open-source LLM was able to extract information from free-text radiology reports with excellent accuracy without requiring further training.

Supplemental material is available for this article.

Scalable and flexible solutions to harness the potential of large unstructured radiology report databases are urgently needed (1). The expected "information revolution" (2) secondary to report structuring has yet to materialize, as large-scale retrospective review of cases of interest still often relies on manual analysis of thousands of free-text reports.

Previous studies have leveraged information contained in radiology reports to estimate the frequency of positive findings to assess the appropriate use of imaging for various indications (3–5). Assessing the relevance of imaging is particularly pertinent for advanced and expansive examinations such as MRI and CT in already high-intensity settings, such as on-call shifts. For instance, the use of MRI for emergency department patients with headache is frequent in Europe and increasing in the United States (6). Ongoing debates over the reported overuse of imaging for this

indication (7) and the potential severity of subtle cases highlight the need for a large volume of data to address the appropriateness of MRI in this indication. However, the post hoc identification of patient cohorts, as well as the extraction of information from free-text reports, rely on time-consuming human reviews of hundreds of reports.

Readily available methods to automate information extraction from imaging reports to address this question are lacking. Existing solutions rely on either *(a)* rule-based methods requiring language- and institution-specific dictionaries or *(b)* bidirectional encoder representations from transformers (or BERT)–based models dependent on subsequent fine-tuning (8). Although well suited for pretargeted tasks, these methods usually lack the expected flexibility to match the diversity of radiology reports and intended aims; tedious and specific training or adaptation to the institutional data are still necessary (1).

Abbreviations

GPT = generative pre-trained transformers, LLM = large language model

Summary

A local open-source large language model extracted information from real-life free-text radiology reports with excellent accuracy without requiring specific training.

Key Points

- Vicuna (LMSYS Org), an on-premise and open-source large language model, reviewed 2398 emergency brain MRI free-text reports and achieved high performance metrics for detecting the presence of headache in the clinical context (F1 score: 98.0 [95% CI: 97.0, 98.7]), detecting the use of contrast medium injection in the protocol (F1 score: 99.6 [95% CI: 99.0, 99.9]), and study categorization as either normal or abnormal (F1 score: 97.3 [95% CI: 95.4, 98.6]).
- Vicuna performed causal inference between a radiologic finding and a symptom with 82% accuracy, a task challenging even for radiologists (interrater κ < 0.80).

Keywords

Large Language Model (LLM), Generative Pretrained Transformers (GPT), Open Source, Information Extraction, Report, Brain, MRI

Large language models (LLMs), owing to their training on extensive natural language databases, exhibit unique adaptability to new tasks in different languages (9). Larger models such as OpenAI's Generative Pre-trained Transformer 4 (GPT-4) and ChatGPT, as well as Google's Bard (now known as Gemini), have demonstrated excellent performance in various medical tasks without requiring fine-tuning (9). However, their cloud-based nature conflicts with the necessity for data confidentiality, leading to debates regarding their compatibility with real-life medical data (10). A recent study on radiology reports had to create fictitious reports to circumvent this issue (11). Fink and colleagues (12) showed that measurements from previous examinations could be extracted from CT reports with GPT-4, while calling for replication of their findings with privacy-preserving models. In this context, the use of LLMs that are accessible on premises, such as LMSYS Org's Vicuna (13), could represent a promising alternative. A recent study assessed the ability of Vicuna to extract findings from chest radiograph reports from a publicly available database (14). However, to the best of our knowledge, no study has investigated the use of open-source LLMs for information extraction from real-life radiology reports. The objective of this study was to assess the performance of a task-agnostic, on-premise LLM in information extraction tasks from real-life radiology reports, using human review as the reference standard.

Materials and Methods

This retrospective study aimed to evaluate the feasibility of automated information extraction from French free-text radiology reports with an open-source LLM. The data warehouse from which the reports were extracted was approved by the French data protection authority (reference no. 2019–103).

Use of the data for this specific study was approved by the Lille University Hospital institutional review board in June 2023 (EDS2307251350).

Data

Free-text reports were obtained from the health data warehouse of our institution. They were pseudonymized by detecting and removing the place of residence of the patient, their name, and the name of the prescribing physician using eHOP software (Université de Rennes). Eligible reports were brain MRI scans from patients in the emergency department performed from January 2022 to December 2022. Reports were segmented into sections—clinical context, protocol, results, and conclusion—with Python, version 3.10 (Python Software Foundation). The report text itself was left unaltered. Reports were initially written in French and were not translated to English for this study. For reference, texts shown in Figure 1 and Tables 1, S1, and S2 were translated to English by the authors and manually modified for de-identification. Reports were initially written in free text by 22 trainees and 21 board-certified radiologists, who were unaware of this study at the time of reporting.

Variable Extraction

Four clinically pertinent information extraction tasks were defined as follows: *(a)* presence of headache as a symptom from the clinical context, *(b)* presence of contrast medium injection from the protocol, *(c)* classification of the examination as either normal or abnormal based on its conclusion, and *(d)* inference of causality between the findings from the conclusion and the headache presented by the patient. All information was initially reported by the radiologist as free text. Age and sex of the patients were automatically inserted in the reports by IMPAX Ris Qdoc, version 6.1 (AGFA Healthcare) and were subsequently collected with regular expression matching using Python, version 3.10.

Ground Truth

Ground truth by task was defined as follows: For tasks 1 and 2, ratings from two in-training radiologists (B.L.G. and A.L., with up to 3 years of experience in neuroradiology) were used. Headache had to be explicitly mentioned in the report to be considered a positive finding. Symptoms had to be acute, recently exacerbated, persistent despite analgesia, or atypical for the patient. For tasks 3 and 4, ratings from two board-certified neuroradiologists (C.B. and G.K., with 7 years and 13 years of experience, respectively) and the two previously involved trainees were used as ground truth. The objective, as initially declared to the raters, was to determine the frequency of findings responsible for headache in patients in the emergency department who underwent a brain MRI examination in our center. Raters had access only to the section of the report related to the task. All ratings were performed independently, with each rater blinded to the ratings of their colleagues and the Vicuna model. Discrepancies were settled by consensus of the two senior radiologists. For task 4, the possible ratings were responsible for headache or not responsible for headache. Then the

Table 1: Examples of Errors by Task and Self-Explanations from Vicuna

Error	Input	Output	Vicuna Self-Explanation
Task 1			
False positive	Episode of visual blurring with fluctuations in blood pressure (as observed by ambulatory blood pressure monitoring) in a patient with a history of migraine. Suspicion of TIA	Symptoms from INDICATION: Visual blurring/other. Fluctuations in blood pressure/other. Migraine/headache. Suspicion of TIA/other	The patient has an episode of visual blurring with fluctuations in blood pressure, which could be a symptom of a transient ischemic attack (TIA). However, the patient also has a history of migraine, which could also explain the visual blurring and fluctuations in blood pressure.
False negative	Suspicion of cerebral venous thrombosis in a cephalalgic patient with a history of cerebral venous thrombosis	Symptoms from INDICATION: Suspicion of cerebral venous thrombosis/other	I apologize for the confusion in my previous response. Upon reflection, I realize that I did not fully address the indication provided by the doctor. The indication mentions a cephalalgic patient with a history of cerebral venous thrombosis, and the task is to list all the symptoms and determine if they correspond to headache or another symptom. Based on the indication, the symptoms that correspond to headache are: Cephalalgia. The other symptoms mentioned in the indication are: History of cerebral venous thrombosis. It is important to note that the patient's history of cerebral venous thrombosis is not a symptom but rather a medical condition that the patient has already been diagnosed with.
Task 2			
False positive	Axial T1, SWI, FLAIR, and diffusion sequences. Coronal T2 phase-contrast venous sequence	Words mentioning injection: phase-contrast venous sequence Injected: Yes	My apologies, I misunderstood the task. I thought the task was to identify if contrast media injection was mentioned in the text, not if it was performed. I will provide a revised response using the template provided: 'Doctor', 'PROTOCOLE: Axial T1, SWI, FLAIR, and Diffusion sequences. Coronal T2 Phase-Contrast Venous sequence.,' 'Robot',-Mention of injection: Phase-Contrast Venous sequence -Injected?: No, I hope this revised response is more accurate. Please let me know if you have any further questions or if there is anything else I can assist you with.

(Table 1 continues)

Table 1 (continued): Examples of Errors by Task and Self-Explanations from Vicuna

Error	Input	Output	Vicuna Self-Explanation
False negative	Axial FLAIR, T2*, and diffusion sequences; ADC mapping. Gado MRA with 3D reconstructions. MRA of the supra-aortic trunks and venous MR angiography sequence	Words mentioning injection: None Injected: No	I apologize for the confusion in my previous response. Upon further analysis, I did not find any explicit mention of injection in the provided protocol. Therefore, I concluded that no injection was performed. However, I understand that the protocol includes Gado MRA with 3D reconstructions and MRA of the supra-aortic trunks and Venous MR Angiography sequence, which may suggest the use of contrast media. I apologize for any confusion caused by my previous response, and I will strive to provide more accurate information in the future.
Task 3			
False positive	No evidence of recent hemorrhagic or ischemic lesions. No intracranial mass effect. No cerebral venous thrombosis. Thin appearance of the right transverse sinus	List of findings: No evidence of recent hemorrhagic or ischemic lesions/ normal. No intracranial mass effect/ normal. No cerebral venous thrombosis/normal. Thin appearance of the right transverse sinus/abnormal	The conclusion states that there is no evidence of recent hemorrhagic or ischemic lesions, no intracranial mass effect, and no cerebral venous thrombosis. Additionally, the right transverse sinus appears to be thin, which could be a cause for concern. It is important to note that the thin appearance of the sinus may be a normal variation or may be related to other factors, so further investigation may be necessary.
False negative	No evidence of recent hemorrhagic or ischemic lesions. No evidence of worsening neurosarcoidosis symptoms. No cerebral venous thrombosis	List of findings: No evidence of recent hemorrhagic or ischemic lesions/ normal. No evidence of worsening neurosarcoidosis symptoms/normal. No cerebral venous thrombosis/ normal.	1. No evidence of recent hemorrhagic or ischemic lesions-This is a normal finding, as it indicates that there are no recent lesions or damage to the brain tissue. 2. No evidence of worsening neurosarcoidosis symptoms-This is a normal finding, as it suggests that there is no significant change or worsening of the symptoms related to neurosarcoidosis. 3. No cerebral venous thrombosis-This is a normal finding, as it indicates that there is no evidence of blood clots in the cerebral veins. Overall, the findings suggest that there are no significant abnormalities or changes in the brain that could explain the patient's symptoms. The stability of the intracranial pressure signs and the absence of recent hemorrhagic or ischemic lesions, worsening neurosarcoidosis symptoms, and cerebral venous thrombosis are reassuring findings.

(Table 1 continues)

Table 1 *(continued)*: Examples of Errors by Task and Self-Explanations from Vicuna

Error	Input	Output	Vicuna Self-Explanation
Task 4			
False positive	Absence of recent ischemic or hemorrhagic lesion. Absence of signs of recent cerebral venous thrombosis. Sequelae of old cerebral venous thrombosis in the left lateral sinus and upper part of the left internal jugular vein	List of findings: Absence of recent ischemic or hemorrhagic lesion/does not usually cause headache. Absence of signs of recent cerebral venous thrombosis/does not usually cause headache. Sequelae of old cerebral venous thrombosis in the left lateral sinus and upper part of the left internal jugular vein/can cause headache	Absence of recent ischemic or hemorrhagic lesion: This means that there is no recent evidence of damage to the brain tissue due to a lack of blood flow or bleeding. This is generally not a cause of headache.
			Absence of signs of recent cerebral venous thrombosis: This refers to the absence of any signs of a blood clot in the veins of the brain. This is also generally not a cause of headache.
			Sequelae of old cerebral venous thrombosis in the left lateral sinus and upper part of the left internal jugular vein: This refers to the residual effects of a previous blood clot in the veins of the brain, specifically in the left lateral sinus and upper part of the left internal jugular vein. This can sometimes cause headache, but it depends on the severity and location of the clot.
False negative	As compared with previous MRI: decreased thickness, stable extent, and expected evolution of the signal from the right petrous carotid hematoma, nonstenosing. Permeability with consistent caliber of the arteries in the circle of Willis	List of findings: As compared with previous MRI: decreased thickness, stable extent, and expected evolution of the signal from the right petrous carotid hematoma, nonstenosing/does not usually cause headache. Permeability with consistent caliber of the arteries in the circle of Willis/does not usually cause headache	As compared with previous MRI: decreased thickness, stable extent, and expected evolution of the signal from the right petrous carotid hematoma, nonstenosing-This finding refers to the comparison of the current MRI results with those from a previous MRI scan. The 'decreased thickness' and 'stable extent' likely indicate that the hematoma has decreased in size and is not growing, while the 'non-stenosing' part suggests that there is no narrowing or blockage in the carotid artery. These findings do not usually cause headaches.
			Permeability with consistent caliber of the arteries in the circle of Willis-This finding refers to the blood flow and the size of the arteries in the circle of Willis, which is a ring of blood vessels at the base of the brain. 'Permeability' likely means that the blood vessels are allowing proper blood flow, and the 'consistent caliber' suggests that the size of the arteries is normal. These findings do not usually cause headaches.
			Absence of cerebral venous thrombosis-This finding indicates that there is no blood clot in the veins of the brain, which can cause headaches.

Note.—ADC = apparent diffusion coefficient, FLAIR = fluid-attenuated inversion recovery, MRA = magnetic resonance angiography, SWI = susceptibility-weighted imaging, TIA = transient ischemic attack.

Example Texts from Dialogues with Vicuna

		Task 1: Symptom of Interest (Headache)
Prompt	Doctor	I will present you short indications for radiology examinations. Your task is to list all the symptoms from them and tell me if they correspond to headache or to another symptom. You will answer only according to this template: Symptoms from INDICATION: -List all the symptoms/is it headache or another symptom ? (If you find no symptom, answer "N/A")
Example 1	Doctor	INDICATION: Headaches and progressively worsening vertigo for the past 2 days, associated with a sensation of hypoesthesia on the right side of the body. Suspicion of cerebral venous thrombosis.
	Robot	Symptoms from INDICATION: -Headaches/headache -Progressively worsening vertigo/other -Sensation of hypoesthesia on the right side of the body/other
Example 2	Doctor	INDICATION: Suspicion of cerebral venous thrombosis.
	Robot	Symptoms from INDICATION:-N/A
Example 3	Doctor	INDICATION: Investigation for intracranial hypertension in a 27-year-old female patient who presented with intermittent visual blur and sudden loss of visual acuity. Associated with bilateral and symmetric papilledema revealed on fundoscopy.
	Robot	Symptoms from INDICATION: -Intermittent visual blur/other -Sudden loss of visual acuity/other -Bilateral and symmetric papilledema revealed on fundoscopy/other
		Task 2: Contrast medium injection
Prompt	Doctor	I will present you protocols from radiology reports. I want to you to tell me if contrast medium injection was performed (often with gadolinium or gado). Ignore all other information, they are insignificant. One crucial rule is that you respond only according to this template: -Words mentioning injection: [retrieve the words mentioning injection or gadolinium or gado in the text] -Injected ?: [Yes/No]
Case 1	Doctor	ID:98678062.0 Axial FLAIR, T2*, SWI, and Diffusion sequences. ADC mapping. Axial 3D-SET1 BV sequence on the cervical carotid and vertebral arteries. Gado-MR with 3D reconstructions./Ms-born: [JJ]/[MM]/1970 (Age: 51)
	Robot	-Words mentioning injection: Gado-MR with 3D reconstructions. -Injected ?: Yes
Case 2	Doctor	ID:13009397.0 Axial Diffusion and FLAIR sequences. Volumetric SWI and 3D TOF sequences./Mr born: [JJ]/[MM]/1988 (Age: 33)
	Robot	-Words mentioning injection: None -Injected ?: No
		Task 3: Classification of the examination as normal or abnormal
Prompt	Doctor	I will present you conclusions from radiology examinations. Your task is to list all the findings from them and tell me if they correspond to normal or abnormal findings. Keep in mind, an abnormal finding may be described as stable or regressing, so please be sure not to miss any abnormal finding. Each conclusion is independent and should not influence your answer. You will answer only according to this template: List of findings: -List all the findings/normal or/abnormal finding ?
Case 1	Doctor	Conclusion: No evidence of recent hemorrhage or ischemic lesions. No arterial occlusion on the Willis polygon. Known right deep hematoma sequela without significant changes.
	Robot	List of findings: -No evidence of recent hemorrhage or ischemic lesions/normal -No arterial occlusion on the Willis polygon/normal -Known right deep hematoma sequela without significant changes/abnormal
Case 2	Doctor	Conclusion: No evidence supporting cervical arterial dissection, RCVS, or cerebral venous thrombosis.
	Robot	List of findings:-No evidence supporting cervical arterial dissection/normal -No evidence supporting RCVS/normal -No evidence of cerebral venous thrombosis/normal
		Task 4: Causal inference between a finding and a symptom (headache)
Prompt	Doctor	I will present you abnormal findings from radiology examinations. Your task is to list all the findings from them and tell me if they usually cause headache or not. Each conclusion is independent and should not influence your answer. You will answer only according to this template: List of findings: -List all the findings/can cause headache or not ?
Case 1	Doctor	Conclusion: No evidence of recent hemorrhagic or ischemic lesions. No arterial occlusion on the Willis polygon. Known right deep hematoma sequela without significant changes.
	Robot	List of findings: -No evidence of recent hemorrhagic or ischemic lesions/does not usually cause headache -No arterial occlusion on the Willis polygon/does not usually cause headache -Known right deep hematoma sequela without significant changes/does not usually cause headache
Case 2	Doctor	Conclusion: Presence of 4 nodular cerebral lesions with ring enhancement suggestive of secondary lesions. A left parieto-occipital parenchymal hematoma probably secondary to a hemorrhagic metastatic lesion.
	Robot	List of findings: -Presence of 4 nodular cerebral lesions with ring enhancement suggestive of secondary lesions/can cause headache -A left parieto-occipital parenchymal hematoma probably secondary to a hemorrhagic metastatic lesion/can cause headache
Case 3	Doctor	Conclusion: No evidence of recent ischemic or hemorrhagic lesions. No MRI findings supporting cervical arterial dissection. Incidental discovery of a 2.5 mm intracranial arterial addition image on the left sylvian region, requiring further evaluation by a specialist.
	Robot	List of findings: -No evidence of recent ischemic or hemorrhagic lesions/does not usually cause headache -No MRI findings supporting cervical arterial dissection/does not usually cause headache -Incidental discovery of a 2.5 mm intracranial arterial addition image on the right sylvian region, requiring further evaluation by a specialist/does not usually cause headache
Case 4	Doctor	Conclusion: Infiltrating lesion with contrast enhancement in the left parietal region, extending to the left insula and crossing the midline to the right occipital region, spanning approximately 6 cm, compatible with an intracerebral lymphoma. Mass effect with early anterior subfalcial herniation observed. Two recent ischemic spots detected (left insular and left cerebellar regions). No proximal occlusion of the arteries in the Willis polygon. A neurosurgical evaluation is recommended.
	Robot	List of findings: -Infiltrating lesion with contrast enhancement in the left parietal region, extending to the left insula and crossing the midline to the right occipital region, spanning approximately 6 cm, compatible with an intracerebral lymphoma/can cause headache -Mass effect with early anterior subfalcial herniation observed/can cause headache -Two recent ischemic spots detected (left insular and left cerebellar regions)/can cause headache -No proximal occlusion of the arteries in the Willis polygon/does not usually cause headache -A neurosurgical evaluation is recommended/does not usually cause headache
Case 5	Doctor	Conclusion: Vascular leukoencephalopathy, chronic microbleeds, and a right thalamic lacuna suggestive of small vessel disease.
	Robot	List of findings: -Vascular leukoencephalopathy/does not usually cause headache -Chronic microbleeds/does not usually cause headache -Right thalamic lacuna/does not usually cause headache

Note.—ADC = apparent diffusion coefficient, FLAIR = fluid-attenuated inversion recovery, SWI = susceptibility-weighted imaging, RCVS = reversible cerebral vasoconstriction syndrome.

Figure 1: Example texts from dialogues with Vicuna 13-B (LMSYS Org).

a single diagnostic group (eg, sinus pathologies) can be considered headache-causing or not depending on the size, shape, signal, enhancement, and adjectives used by the reporting neuroradiologist (as well as potential changes from previous examinations) (4). Consensual examinations were defined as reports for which at least three radiologists rated the association between the main finding and the headache as certain.

Sample Size Calculation

Sample size was calculated using a previously developed equation (15). To estimate sensitivity and specificity for task 4 of 0.7 (16) with a 95% CI, a precision of 10%, and an expected prevalence of causal findings of 20% (4), the number of reports should be at least 400. With approximately 20% of emergency MRI examinations performed for headache in our institution, at least 2000 reports were screened.

Model

Version 1.3 of the Vicuna 13-B model was used (https://huggingface.co/lmsys/vicuna-13b-v1.3). Vicuna is based on the LLaMA model by Meta and fine-tuned on conversations shared by users on ShareGPT (12). We made no alterations to the model. To mitigate variability, temperature was set to 0. Interaction with the model was made through FastChat (https://github.com/lm-sys/FastChat) (17). We developed a Python script to automate interactions with Vicuna and to control the quality of its output (https://github.com/BastienLeGuellec/RadioVicuna). The model had access to the same section of the report as the human raters. The final script took free-text reports as input and provided a table as output (Fig 2). The model was run on two Quadro RTX 6000 graphics processing units (NVIDIA).

Prompting

Prompts were written in English. They were designed to be short and simple to adapt intuitively to new tasks and align with empirical best practices for prompting LLMs (18). Vicuna was specifically prompted to segment its answers to improve the transparency of error analysis (Table 1). We used few-shot in-context learning: fake contextual examples created manually were provided to the model (19). Increasing numbers of examples were used until saturation of diagnostic performance. Examples were engineered to depict the variability of the sample (positive and negative examples, frequent phrasing used by the radiologists). In a sensitivity analysis, we repeated the same tasks with prompts translated from English to French (Tables S3 and S4). For each error displayed in

rater could determine whether they were certain or not certain about their rating. The causal nature of a finding was left to the appreciation and experience of the raters, as done in previous studies (4,5). No predefined list of pathologies was established to account for the specificity of each report, as findings from

Table 2: Vicuna Performance and Calculation Times for Each Extraction Task

Variable	Task 1: Symptom	Task 2: Contrast Material Injection	Task 3: Classification*	Task 4: Causal Inference†
Sensitivity (%)	98.0 (583/595) [96.5, 99.0]	99.4 (514/517) [98.3, 99.9]	96.5 (219/227) [93.2, 98.5]	88.2 (120/136) [81.6, 93.1]
Specificity (%)	99.3 (1791/1803) [98.8, 99.7]	99 (68/69) [92, 100]	98.9 (364/368) [97.2, 99.7]	73 (66/91) [62, 81]
F1 score	98.0 (1166/1190) [97.0, 98.7]	99.6 (1028/1032) [99.0, 99.9]	97.3 (438/450) [95.4, 98.6]	85.4 (240/281) [80.7, 89.3]
Accuracy (%)	99.0 (2374/2398) [98.5, 99.4]	99.3 (584/588) [98.3, 99.8]	98.0 (583/595) [96.5, 99.0]	81.9 (186/227) [76.4, 86.7]
Calculation time for all reports (min)‡	180	45	60	30
Prompt engineering (min)	30	30	45	60

Note.—Data in parentheses are numerators and denominators, with 95% CIs in brackets.
* Normal or abnormal.
† Casual inference between the MRI findings and headache.
‡ The model was run on two Quadro RTX 6000 graphics processing units (NVIDIA).
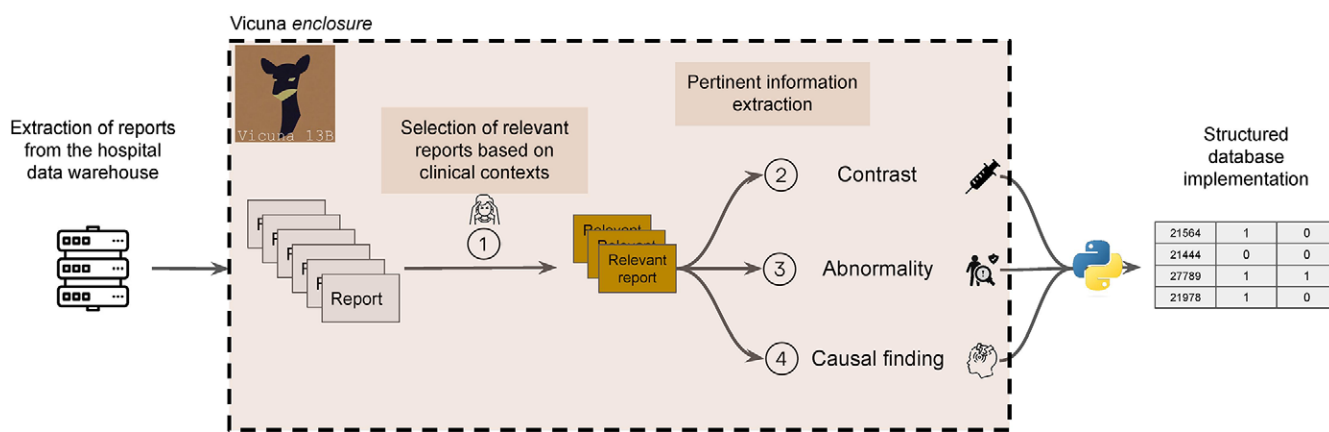
Figure 2: Diagram shows workflow for automated information extraction from pseudonymized radiology reports with Vicuna 13-B (LMSYS Org), an open-source large language model. Four tasks are defined as follows: *1*, reports' triaging based on the presence of headache; *2*, extraction of contrast medium injection from the protocol; *3*, classification of the examination as either normal or abnormal based on the conclusion; and *4*, causal inference between the main finding and headache.

Table 1, we prompted the model to detail its initial rating in a zero-shot approach (no additional example was provided to the model for this prompt): "Reflect on your answer and detail it" (Table 1).

Performance Evaluation and Statistical Analysis

We estimated sensitivity, specificity, accuracy, and F1 scores, with 95% binomial CIs, against the human-established ground truth. Interrater agreement was measured with Fleiss κ statistics. We tested the robustness of the results with different numbers of contextual examples. For task 4, comparison of the model accuracy between the consensual reports and the rest of the reports was performed with a χ^2 test. Comparison of performance between prompts in English and French was tested with a McNemar test. $P < .05$ was considered to indicate statistical significance. All statistical analyses were performed by a statistician (A.H.) not involved in the rating of the reports using R software, version 4.3.1 (R Foundation).

Results

Radiologist Ratings

Among the 2398 brain MRI scans included in the study, radiologists identified 595 scans performed for headache (median age of patients, 35 years [IQR: 26–51 years]; 403 [68%] women and 192 [32%] men). In seven of those reports, contrast medium injection could not be inferred directly from the text, and so those reports were excluded from the task 2 analysis. Of the remaining 588 scans, 517 (88%) were performed with contrast medium injection. Thirty-eight percent (227 of 595) of reports included an abnormal finding in the conclusion (interrater Fleiss κ: 0.96 [95% CI: 0.95, 0.99]). Twenty-three percent (136 of 595) included at least one abnormality that could certainly or probably explain the patients' headaches (interrater Fleiss κ: 0.77 [95% CI: 0.72, 0.83]) (Figs 3 and S1). Sixteen percent (95 of 595) reported unrelated findings that were not the cause of headache. Diagnoses are available in Table S5. The ranges of the estimated
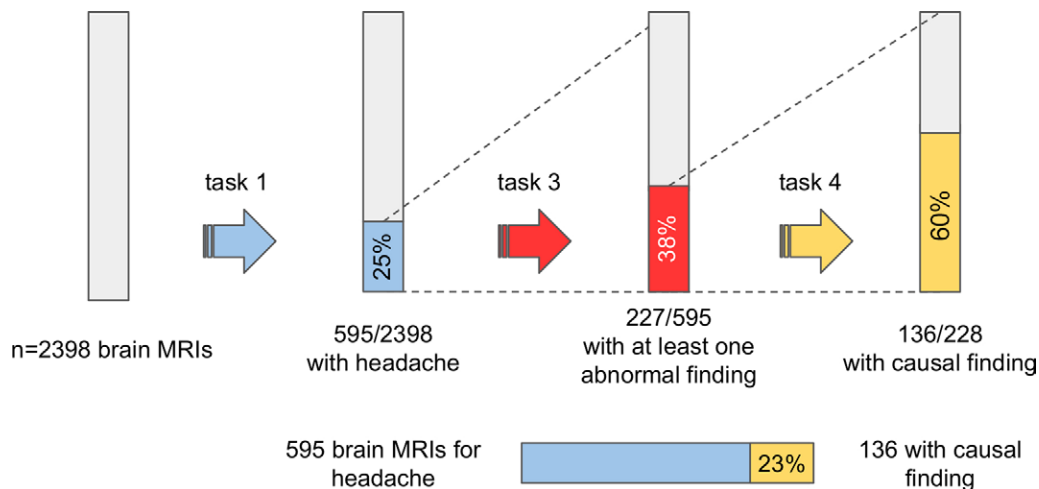
Figure 3: Graph shows diagnostic yield of brain MRI scans in patients in the emergency department with headache.

time spent by each radiologist to review and rate the examinations was 180–210 minutes for the 2398 examinations of task 1, 90–105 minutes for the 588 examinations of task 2, and 100–140 minutes for tasks 3 and 4.

Vicuna Performance

All performance metrics were more than 95% for assessment of the presence of headache as a symptom in the indication (sensitivity, 98.0% [95% CI: 96.5, 99.0]; specificity, 99.3% [95% CI: 98.8, 99.7]; accuracy, 99.0% [95% CI: 98.5, 99.4]), contrast medium injection assessment (sensitivity, 99.4% [95% CI: 98.3, 99.9]; specificity, 98.6% [95% CI: 92.2, 100.0]; accuracy, 99.3% [95% CI: 98.3, 99.8]), and classification of the examination as either normal or abnormal based on the conclusion of the report (sensitivity, 96.0% [95% CI: 92.5, 98.2]; specificity, 98.9% [95% CI: 97.2, 99.7]; accuracy, 97.8% [95% CI: 96.3, 98.8]). Causal inference between the main findings of the examination and the patient's headache as a presenting symptom was accurate in 81.9% (95% CI: 76.4, 86.7) of cases (sensitivity, 88.2% [95% CI: 81.6, 93.1]; specificity, 73.0% [95% CI: 62, 81]) (Table 1 and Fig 4). For this task, Vicuna performed significantly better on consensual examinations (accuracy, 89.7% [95% CI: 82.6, 84.5] on 116 consensual examinations; accuracy, 73.9% [95% CI: 64.7, 81.8] on the remaining 111 reports; P = .002) (Table S6). No change in performance metrics was observed when prompting Vicuna in French (Tables S3 and S4).

Prompts were engineered in 30 minutes for tasks 1 and 2 and 1 hour for tasks 3 and 4. Calculation time on our machine ranged from 30 minutes for task 4 (227 reports) to 3 hours for task 1 (2398 reports) (Table 1). Prompts used and examples of answers are provided in Figure 1. Contingency matrices are available in Table S7.

Providing fake contextual examples (few-shot prompting) systematically improved performance over the zero-shot approach. Saturation of accuracy was obtained with four to six examples, depending on the task (Table S8). Contextual examples are provided in Table S1.

Vicuna Errors Analysis

For all tasks, no systematic error pattern could be identified (Fig 1). False-positive and false-negative errors included various phrasings, contexts, and findings. For task 4, aneurysms were frequently marked as responsible for headache by Vicuna (four false-positive errors), and neither of the two cytotoxic corpus callosum lesions were marked as positive findings (false-negative errors). No other finding was frequently associated with errors. Among the 236 reports on which all four tasks were performed, 49 (22%) had a misclassification for one task, and only one report had a misclassification for two different tasks (Table S9).
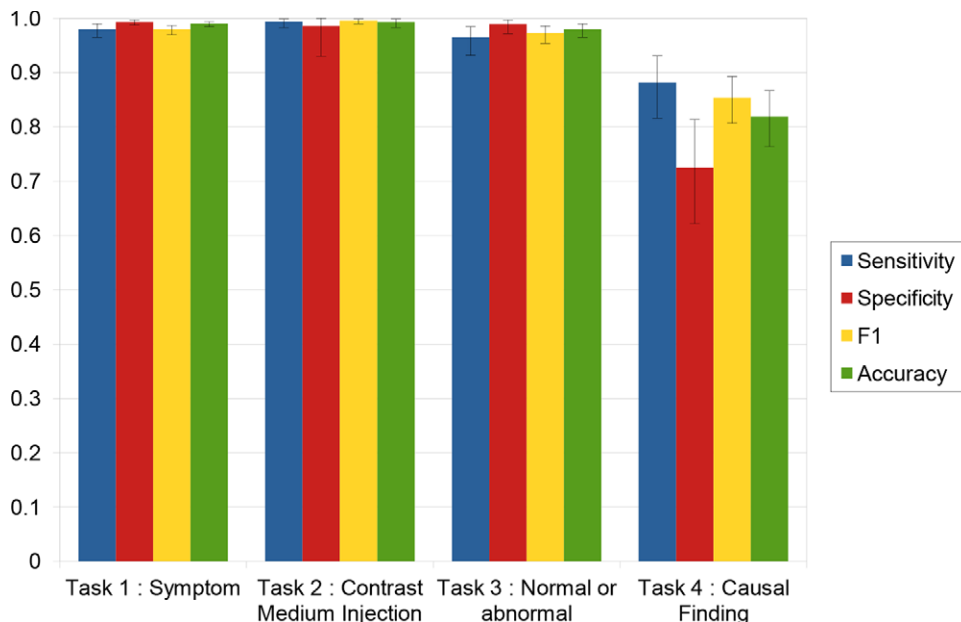
When prompted to self-analyze its responses, Vicuna provided the user with additional information or self-corrections, displayed in Figure 1.

Discussion

An on-premise and open-source LLM extracted information from real-life free-text radiologic reports with excellent accuracy (>95% across three tasks) and performed causal inference between findings and a symptom with 82% accuracy, without requiring specific training.

Human manual review of 2398 brain MRI reports for patients in the emergency department with headache revealed that 25% of emergency MRI examinations were performed for patients with headache, and among those, 23% had significant findings that could explain the headache. A recent meta-analysis (20) reported a rate of causal findings of 9%, significantly lower than ours. That review pooled studies with different imaging modalities (MRI and CT), regions, and center types (secondary, tertiary). Our results are in line with a recent study (4) on brain MRI scans for nontraumatic headache conducted in a Finnish tertiary center, yielding a 20% rate of causal findings.

Performances of commercially available LLMs for feature extraction tasks are still debated (21,22). Initial studies using GPT-3 (OpenAI) on benchmark data reported accuracy below 60% for various name-entity recognition and relation extraction tasks (21). More recent works on GPT-4 reported higher accuracy on simpler tasks, notably more than 90% precision to

Figure 4: Graph shows performance of Vicuna 13-B (LMSYS Org) for tasks 1, 2, 3, and 4. Error bars are 95% CIs.

ensures that the method is easily adaptable for a variety of new tasks across different centers and languages without the need for extensive setup. The model's comprehensive initial training enables it to effectively manage synonyms, negations, and circumlocutions without the need for further postprocessing, which sets it apart from traditional text-searching techniques. The code for our scripts is made available, and a practical guide is provided *(https://github.com/BastienLeGuellec/RadioVicuna)*. Because the model we used is inherently task agnostic, we hypothesize that our method could also be applied to other medical texts beyond radiology reports (29–31).

Explainability of artificial intelligence is a critical aspect of its practical implementation, especially in health care. Our approach allows for direct engagement with the LLM to solicit explanations for each evaluation it provides, ensuring that its answers are based on plausible medical knowledge and correct understanding of the task. Other methods exist to assess the level of certainty for LLM ratings, including examining the probability associated with each output token, a proxy for confidence level (32). Accessing this confidence metric is possible with our proposed method, with detailed instructions available in the practical guide. Future work will need to integrate additional aspects of LLM explainability (33), enhancing transparency and trustworthiness of the proposed workflows.

In the future, LLMs will likely serve as powerful tools for several individuals: *(a)* radiologists, by extracting information from previous examinations to facilitate follow-up (12), drafting reports to improve workflow (34), or checking for readability and completeness of their reports; *(b)* clinicians, by annotating or structuring the report to improve its readability; and *(c)* patients, by providing an assistant able to explain the examination and answer related questions in real time (35). Although recent studies have highlighted the diagnostic abilities of LLMs in the field of radiology (16), LLMs may be integrated sooner into the workflow of radiologists as assistants for low-expertise tasks. This integration is expected to enhance and promote the radiologists' expertise rather than replace it, relieve them of time-consuming tasks, and facilitate more effective communication of information in their reports. Moreover, from a research perspective, LLMs could substantially contribute to addressing the scarcity of high-quality labeled data for training computer vision models. By leveraging the expert information already present in the report, LLMs could enable a substantial increase in the volume of labeled data, with a limited cost.

This study had limitations. First, because this was a single-center study, we could not test for variability in center writing

retrieve medication names from public datasets, improving with the use of self-verification methods (22). Using GPT-4, Adams and colleagues (11) structured fictitious radiology reports with 100% accuracy, and Fink and colleagues (12) successfully extracted measurements from cancer lesions in 96% of reports. Along with these results, we found that a lesser-powered open-source LLM chatbot can extract a symptom from a clinical context and the presence of contrast medium injection from an MRI protocol with near-perfect accuracy. Recent studies emphasize the innate medical knowledge of LLMs and their logical abilities (23,24), which improved with a low number of contextual examples (19). Although the causal inference task described herein relies on the initial inference process of listing findings from images, our results offer insights into a new approach for inferring causality in radiology reports (25).

The ability of open-source models to run on premises has multiple advantages. First, the absence of communication of medical data to private third-party servers ensures critically needed data privacy (10). Second, as opposed to cloud-based methods relying on ever-changing algorithms, reproducibility over time (version control) can be ensured (26). This allows for replicability of results and monitoring model drift during continuous optimization. Third, the training of task-agnostic models on diverse natural texts distinctively separates their intended use as chatbots from their emerging application in medical texts. This separation significantly mitigates the risk of overfitting, as the model's training data do not overlap with the specific data it encounters in clinical settings. Finally, no additional cost is associated with the use of local models. As recently stated (27,28), the cost per token generated for commercial method scaling with the volume of text to analyze may hamper their widespread use.

The method we propose is based on a publicly available task-agnostic model that was not altered for this study. This approach

styles and languages. However, our center is particularly large (165 000 emergency visits annually), and our dataset is diverse with 43 different writers. Because our prompts were in English and reports were in French, we anticipate high robustness to language variations (36), as also evidenced by similar performances after translating the prompts from English to French. Second, the available clinical information was restricted to brief contexts from the report itself. Further studies should extend the scope of this method to data extraction from unsegmented clinical reports. Third, the reference standard for the most complex task of causal inference between a finding and a symptom is in itself subjective, based on the experience of the radiologists. As in previous studies, ground truth was defined by expert consensus (4,5). The difficulty of this task, as shown by the high rate of nonconsensual examinations in our studies (49%), may originate from the frequent lack of explicit comment by the reporting neuroradiologist on the significance of findings. Future works on LLMs, which have shown promise in detailing and simplifying radiology reports (35), may address this issue. Finally, we used a model that will potentially be outperformed by new developments, especially with the recent release of LLaMA 2 (37). However, this workflow is not limited to a specific model or set of models, allowing for flexibility with newer technologies. Further work is needed to study the respective strengths and limits of newer available models.

In conclusion, this proof-of-concept study shows the potential of open-source LLMs to perform information extraction tasks from real-life free-text radiology reports, without the need for additional training. Given the rising social and scientific need for transparency in artificial intelligence, the open-source framework we provide creates a stepping stone for replication studies and may open a new horizon for automated analysis of unstructured medical texts.

References

1. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural Language Processing in Radiology: A Systematic Review. Radiology 2016;279(2):329–343.
2. Langlotz CP. Automatic structuring of radiology reports: harbinger of a second information revolution in radiology. Radiology 2002;224(1):5–7.
3. Lungren MP, Amrhein TJ, Paxton BE, et al. Physician self-referral: frequency of negative findings at MR imaging of the knee as a marker of appropriate utilization. Radiology 2013;269(3):810–815.
4. Happonen T, Nyman M, Ylikotila P, Merisaari H, Mattila K, Hirvonen J. Diagnostic yield of emergency MRI in non-traumatic headache. Neuroradiology 2023;65(1):89–96.
5. Budweg J, Sprenger T, De Vere-Tyndall A, Hagenkord A, Stippich C, Berger CT. Factors associated with significant MRI findings in medical walk-in patients with acute headache. Swiss Med Wkly 2016;146:w14349.
6. Trofimova AV, Duszak R Jr, Kadom N, Sadigh G. Increasing and disparate use of neuroimaging for adults and children with non-traumatic headaches in the US emergency departments: Opportunities for improvement. Headache 2021;61(1):179–189.
7. Dunne CL, Elzinga JL, Vorobeichik A, et al. A Systematic Review of Interventions to Reduce Computed Tomography Usage in the Emergency Department. Ann Emerg Med 2022;80(6):548–560.
8. Jupin-Delevaux É, Djahnine A, Talbot F, et al. BERT-based natural language processing analysis of French CT reports: Application to the measurement of the positivity rate for pulmonary embolism. Res Diagn Interv Imaging 2023;6:100027.
9. Lee P, Bubeck S, Petro J. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. N Engl J Med 2023;388(13):1233–1239.
10. Marks M, Haupt CEAI. AI Chatbots, Health Privacy, and Challenges to HIPAA Compliance. JAMA 2023;330(4):309–310.
11. 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(4):e230725.
12. Fink MA, Bischoff A, Fink CA, et al. Potential of ChatGPT and GPT-4 for Data Mining of Free-Text CT Reports on Lung Cancer. Radiology 2023;308(3):e231362.
13. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. LMSYS Org. https://lmsys.org/blog/2023-03-30-vicuna. Accessed July 9, 2023.
14. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of Using the Privacy-preserving Large Language Model Vicuna for Labeling Radiology Reports. Radiology 2023;309(1):e231147.
15. Buderer NMF. Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. Acad Emerg Med 1996;3(9):895–900.
16. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. Radiology 2023;307(5):e230582.
17. Zheng L, Chiang WL, Sheng Y, et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. ArXiv 2306.05685 [preprint] https://arxiv.org/abs/2306.05685. Posted June 9, 2023. Accessed May 29, 2024.
18. Weng L. Prompt Engineering. https://lilianweng.github.io/posts/2023-03-15-prompt-engineering/. Published 2023. Accessed November 21, 2023.
19. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large Language Models are Few-Shot Clinical Information Extractors. ArXiv 2205.12689 [preprint] https://arxiv.org/abs/2205.12689. Posted May 25, 2022. Accessed May 29, 2024.
20. Jang YE, Cho EY, Choi HY, Kim SM, Park HY. Diagnostic Neuroimaging in Headache Patients: A Systematic Review and Meta-Analysis. Psychiatry Investig 2019;16(6):407–417.
21. Gutiérrez BJ, McNeal N, Washington C, et al. Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again. ArXiv 2203.08410 [preprint] https://arxiv.org/abs/2203.08410. Posted March 16, 2022. Accessed May 29, 2024.
22. Gero Z, Singh C, Cheng H, et al. Self-Verification Improves Few-Shot Clinical Information Extraction. ArXiv 2306.00024 [preprint] https://arxiv.org/abs/2306.00024. Posted May 30, 2023. Accessed May 29, 2024.
23. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023;620(7972):172–180. [Published correction appears in Nature 2023;620(7973):E19.]
24. Webb T, Holyoak KJ, Lu H. Emergent analogical reasoning in large language models. Nat Hum Behav 2023;7(9):1526–1541.
25. Sebro RA, Kahn CE Jr. Automated detection of causal relationships among diseases and imaging findings in textual radiology reports. J Am Med Inform Assoc 2023;30(10):1701–1706.
26. Spirling A. Why open-source generative AI models are an ethical way forward for science. Nature 2023;616(7957):413.
27. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. Lancet Digit Health 2023;5(6):e333–e335.
28. López-Úbeda P, Martín-Noguerol T, Luna A. Radiology in the era of large language models: the near and the dark side of the moon. Eur Radiol 2023;33(12):9455–9457.
29. Kulkarni PA, Singh H. Artificial Intelligence in Clinical Diagnosis: Opportunities, Challenges, and Hype. JAMA 2023;330(4):317–318.
30. Uddin Y, Nair A, Shariq S, Hannan SH. Transforming primary healthcare through natural language processing and big data analytics. BMJ 2023;381:948.
31. Zhang H, Jethani N, Jones S, et al. Evaluating Large Language Models in Extracting Cognitive Exam Dates and Scores. medRxiv 2023.07.10.23292373 [preprint] 2023.07.10.23292373. https://www.medrxiv.org/con-

tent/10.1101/2023.07.10.23292373. Posted February 13, 2024. Accessed May 29, 2024.

32. Jiang Z, Araki J, Ding H, Neubig G. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. Trans Assoc Comput Linguist 2021;9:962–977.

33. Zhao H, Chen H, Yang F, et al. Explainability for Large Language Models: A Survey. ACM Trans Intell Syst Technol 2024;15(2):1–38.

34. Langlotz CP. The Future of AI and Informatics in Radiology: 10 Predictions. Radiology 2023;309(1):e231114.

35. Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for Simplifying Radiology Reports. Radiology 2023;309(2):e232561.

36. Lai VD, Ngo NT, Ben Veyseh AP, et al. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. ArXiv 2304.05613 [preprint] https://arxiv.org/abs/2304.05613. Posted April 12, 2023. Accessed May 29, 2024.

37. Llama 2. Meta AI. https://ai.meta.com/llama/. Accessed July 31, 2023.