

# Boosting LLM-Assisted Diagnosis: 10-Minute LLM Tutorial Elevates Radiology Residents' Performance in Brain MRI Interpretation

Su Hwan Kim <sup>1</sup>, Severin Schramm <sup>1</sup>, Jonas Wihl <sup>1</sup>, Philipp Raffler <sup>1</sup>, Marlene Tahedl <sup>1</sup>, Julian Canisius <sup>1</sup>, Ina Luiken <sup>2</sup>, Lukas Endrös <sup>2</sup>, Stefan Reischl <sup>2</sup>, Alexander Marka <sup>2</sup>, Robert Walter <sup>2</sup>, Mathias Schillmaier <sup>2</sup>, Claus Zimmer <sup>1</sup>, Benedikt Wiestler\* <sup>1</sup>, Dennis M. Hedderich\* <sup>1</sup>

<sup>1</sup>*Department of Diagnostic and Interventional Neuroradiology, Klinikum rechts der Isar, School of Medicine and Health, Technical University of Munich, Munich, Germany*

<sup>2</sup>*Department of Diagnostic and Interventional Radiology, Klinikum rechts der Isar, School of Medicine and Health, Technical University of Munich, Munich, Germany*

\* *These authors contributed equally.*

## Abstract

### Purpose

To evaluate the impact of a structured tutorial on the use of a large language model (LLM)-based search engine on radiology residents' performance in LLM-assisted brain MRI differential diagnosis.

### Materials & Methods

In this retrospective study, nine radiology residents determined the three most likely differential diagnoses for three sets of ten brain MRI cases with a challenging yet definite diagnosis. Each set of cases was assessed 1) with the support of conventional internet search, 2) using an LLM-based search engine (© Perplexity AI) without prior training, or 3) with LLM assistance after a structured 10-minute tutorial on how to effectively use the tool for differential diagnosis. The tutorial content was based on the results of two studies on LLM-assisted radiological diagnosis and included a prompt template. Reader responses were rated using a binary and numeric scoring system. Reading times were tracked and confidence levels were recorded on a 5-point Likert scale. Binary and numeric scores were analyzed using chi-square tests and pairwise Mann-Whitney U tests each. Search engine logs were examined to quantify user interaction metrics, and to identify hallucinations and misinterpretations in LLM responses.

## **Results**

Radiology residents achieved the highest accuracy when employing the LLM-based search engine following the tutorial, indicating the correct diagnosis among the top three differential diagnoses in 62.5% of cases (55/88). This was followed by the LLM-assisted workflow before the tutorial (44.8%; 39/87) and the conventional internet search workflow (32.2%; 28/87). The LLM tutorial led to significantly higher performance (binary scores:  $p = 0.042$ , numeric scores:  $p = 0.016$ ) and confidence ( $p = 0.006$ ) but resulted in no relevant differences in reading times. Hallucinations were found in 5.1% of LLM queries.

## **Conclusion**

A structured 10-minute LLM tutorial increased performance and confidence levels in LLM-assisted brain MRI differential diagnosis among radiology residents.

## **Clinical Relevance Statement**

Our findings highlight the considerable benefits that even low-cost, low-effort educational interventions on LLMs can provide. Integrating LLM education in radiology training programs could augment practitioners' capacity to harness AI technologies effectively.

## Introduction

Large language models (LLMs) are advanced artificial intelligence (AI) systems capable of processing and generating human language. Trained on vast amounts of text data and based on an innovative transformer architecture, these models have demonstrated remarkable performance in various tasks across sectors (1).

With the rapid technological advancements of LLMs in recent years, numerous studies have explored applications of LLMs in radiological workflows. These include the definition of imaging protocols (2–4), performing differential diagnosis based on case presentations (5–9), error checking in radiology reports (10), generation of impressions in radiology reports (11,12), information extraction from free-text radiology reports (13–15) and more. Yet, despite the promising applications, the integration and adoption of LLMs in radiology is not without challenges. Data privacy concerns, bias and error propagation, lack of contextual understanding, and overreliance have been pointed out as relevant limitations of LLMs (1,16–19). Against this background, the critical role of educating healthcare professionals on the appropriate use and potential pitfalls of LLMs has been emphasized (20–22). This may include training in prompt engineering, which describes the strategic crafting of a textual instruction that serves as input to LLMs (22).

One area where insufficient human oversight of LLMs could lead to clinical errors is radiological differential diagnosis. An earlier study on human-LLM collaboration in brain MRI differential diagnosis found that inadequate formulation of prompts can result in misleading LLM outputs, and lacking critical validation of LLM responses can lead to incorrect conclusions (23).

However, whether and how radiology readers can be trained to more effectively apply LLMs in radiological differential diagnosis has not been investigated yet. This study

therefore aimed to evaluate the impact of a structured LLM tutorial on the performance of radiology residents in brain MRI differential diagnosis.

## Methods

Informed patient consent was waived by the Ethics Committee of the Technical University of Munich.

### Study Sample

Thirty challenging brain MRI exams acquired between 01/01/2016 and 12/31/2023 were selected from the local Picture Archiving and Communication System (PACS) system and randomized into three sets (Figure 1). Included exams were deemed as sufficiently complex for use in radiological board certification exams by two board-certified neuroradiologists (DMH and BW) and contained an abnormal finding with a confirmed diagnosis (histopathologically or through independent agreement of at least two neuroradiologists). In each exam, one or more arrows marked the image finding in question. The included exams have been published previously (24). A case overview is provided in Supplement 1.

Nine radiology residents with less than six months of neuroradiology experience were recruited from the local departments of radiology and neuroradiology and randomized into three groups (Table 1). Informed consent was provided by all participants.

### Study Design

Over the course of three sessions, each reader assessed three sets of ten brain MRI cases with varying workflows and provided up to three differential diagnoses for the annotated image findings, ranked by likelihood. Each case was reviewed only once per reader. To control for the confounding effects of case difficulty, each set of cases was assessed by the same number of readers with each workflow (Figure 1). For

every case, demographic and condensed medical history was provided. Sessions took place between 01/03/2024 and 22/05/2024.

First, conventional internet research was conducted to support differential diagnosis, either using web-based search engines, e.g. Google Search, or directly accessing trusted websites (Conventional). Residents were instructed to behave as they would in a clinical routine setting, mimicking the current practice in clinical care. Second, readers utilized an LLM-based search engine (© Perplexity AI Inc., San Francisco, USA) but didn't receive any training beforehand (LLM-Pre-Training). PerplexityAI had been chosen as LLM interface for its ability to access real-time web content and provide source citations. Search queries were powered by GPT-4-Turbo (Generative Pre-trained Transformer 4 Turbo) by OpenAI. Third, another subset of ten cases was evaluated with the assistance of PerplexityAI. This time, however, the session was preceded by a structured tutorial on how to effectively use the tool (LLM-Post-Training). Tutorial details are provided below. In both LLM-assisted workflows, participants were allowed to conduct additional internet search to validate LLM suggestions.

Reading times were recorded using a time tracking software (Toggl Track, © Toggl OÜ, Tallinn, Estonia). Confidence levels were documented for each case on a 5-point Likert scale (1: not at all confident, 5: very confident). Following the second and third session, readers completed questionnaires to evaluate the experience with the LLM-assisted workflow.

## LLM Tutorial

In a short tutorial of no more than 10 minutes, readers were given tips on how to effectively utilize the LLM-based search engine. The content of the tutorial was based on two earlier studies on the application of LLMs for brain MRI differential diagnosis.

One study evaluated the contribution of varying multimodal input elements on the diagnostic performance of GPT4(V) and identified the textual description of radiological image findings as the key element (24). The other demonstrated superior accuracy of LLM-assisted differential diagnosis over a workflow supported by a conventional search engine, but also determined several pitfalls in human-LLM interaction (23). The full script of the tutorial is provided below:

*“A detailed description of image findings is by far the most important factor for accurate LLM responses. The description should include details about location, contrast enhancement, morphology, size and more. An accurate description of the finding location is particularly critical, an inaccurate specification of the location can result in misleading suggestions.*

*Providing relevant information about the medical history can improve the accuracy of LLM responses. However, clinical information unrelated to the image finding might result in misleading LLM outputs. Therefore only clinical information deemed to be relevant for the image finding in question should be provided.*

*Uploading screenshots of key image findings can help improve LLM responses, although their effect is only marginal.*

*Use of connotative terminology can lead to bias and should be avoided (e.g. the term ‘juxtacortical’ is strongly associated with multiple sclerosis).*

*Instructions can be made regarding the extent (number of differential diagnoses mentioned) and format (bullet points, table) of the LLM output.”*

Readers were further encouraged to use the following prompt template:



*“You are a senior neuroradiologist. Below, you will find information regarding a brain MRI scan. Based on this information, identify the three most likely differential diagnoses, ranked by their likelihood. Present your findings in a table format with the following columns: ‘Rank’, ‘Differential Diagnosis’, and ‘Explanation’.*

*[Medical history]*

*[Image description]”*

## Analysis

To ensure performance did not solely depend on prior knowledge but the quality of research, cases where the correct diagnosis could be determined confidently without further research were excluded.

Accuracy of differential diagnoses was determined using two different scoring systems, as described previously (23). The first method used a binary scoring system, where responses were labeled as "correct" if the correct diagnosis was included among the submitted differentials, and "incorrect" if it was not. The second approach assigned scores ranging from 0 to 3 based on the rank of the correct diagnosis within the response (0: correct diagnosis not included, 1: correct diagnosis ranked third, 2: correct diagnosis ranked second, 3: correct diagnosis ranked first). Cases where a correct but less granular response was indicated were rated in consensus (by SHK and SS). For binary scores, a chi-square test was initially applied across all groups, followed by pairwise chi-square tests. For numeric scores and confidence levels, a Kruskal-Wallis test was used to assess differences among all groups, with subsequent pairwise comparisons conducted using the Mann-Whitney U test. To control for false discovery rates, p-values were adjusted using the Benjamini-Hochberg procedure for both scores and confidence. Reading times were analyzed using an ANOVA test

across all workflows, followed by pairwise t-tests. The significance level was set at  $p < 0.05$ . 5-point Likert-scale questionnaire results are reported using descriptive statistics.

Logs of the LLM-based search engine (Perplexity AI) were examined to quantify the number of queries and source references. Queries were categorized by query type (keyword-based vs instruction-based) and by content (differential diagnosis, radiographic features, sample images, anatomy, other). Sources were classified into journal articles and other online sources. The content of LLM responses were screened for incorrect or inconsistent information by two radiology residents (SHK and SS; 1.5 and 2.5 years of experience in reading brain MRI exams) and confirmed by a certified neuroradiologist (DH). As described previously (25), incorrect responses were classified into hallucinations (inconsistent with widely accepted radiological knowledge), misinterpretations (miscomprehending a question and giving contextually irrelevant replies), and clarifications (lacking comprehension of a prompt requiring its rephrasing). Radiographic features described in LLM responses were checked against reference articles of [www.radiopaedia.org](http://www.radiopaedia.org), which is a validated source of radiological knowledge.

Data curation, analysis and visualization were performed using Python (version 3.9.7).

## Results

8 out of 270 responses (3.0%) were excluded from the analysis as readers were able to determine the correct diagnosis confidently without requiring further research. A sample LLM query and its results are shown in Figure 2. 12 out of 262 cases (4.6%) required a consensus decision because the reader provided a correct but less specific diagnosis (e.g. “encephalitis” was counted as correct in a case of limbic encephalitis).

### Binary and Numeric Scores

Based on the binary scoring system, 62.5% (55/88) of responses in the LLM-Post-Training workflow were correct, compared to 44.8% (39/87) in the LLM-Pre-Training and 32.2% (28/87) in the Conventional group (Figure 3). An initial chi-square test across all groups indicated a significant overall difference ( $p < 0.001$ ). Subsequent pairwise chi-square tests showed significant differences between LLM-Pre-Training and LLM-Post-Training ( $p = 0.042$ ), as well as between LLM-Post-Training and Conventional ( $p < 0.001$ ), but not between LLM-Pre-Training and Conventional ( $p = 0.119$ ) (Table 2).

Comparison of numeric scores revealed a median score of 3 in the LLM-Post-Training group, compared to 0 for both the LLM-Pre-Training and Conventional groups (Figure 3). The Kruskal-Wallis test confirmed significant overall differences among the workflows ( $p < 0.001$ ). Pairwise comparisons using the Mann-Whitney U test revealed significant differences between LLM-Pre-Training and LLM-Post-Training ( $p = 0.016$ ) and between LLM-Post-Training and Conventional ( $p < 0.001$ ), but not between LLM-Pre-Training and Conventional ( $p = 0.092$ ) (Table 3).

## Confidence

Median confidence ratings were highest in LLM-Post-Training (median = 4), followed by LLM-Pre-Training (median = 3) and Conventional (median = 3). The proportion of high or very high confidence ratings (4 or 5) was 18% for Conventional, 31% for LLM-Pre-Training, and 54% for LLM-Post-Training (Figure 4). The Kruskal-Wallis test showed a significant overall difference in confidence ( $H = 27.20$ ,  $p < 0.001$ ). Pairwise Mann-Whitney U tests revealed a highly significant difference between Conventional and LLM-Pre-Training ( $p = 0.006$ ), LLM-Pre-Training and LLM-Post-Training ( $p = 0.006$ ) as well as between Conventional and LLM-Post-Training ( $p < 0.001$ ).

## Reading Times

Mean reading times amounted to 07:43 min (Conventional), 08:59 min (LLM-Pre-Training) and 08:35 min (LLM-Post-Training) (Figure 5). An ANOVA test showed a statistically significant overall difference ( $p = 0.030$ ). Pairwise t-tests showed significant differences between Conventional and LLM-Pre-Training ( $p = 0.013$ ), but not between LLM-Pre-Training and LLM-Post-Training ( $p = 0.403$ ) or between Conventional and LLM-Post-Training ( $p = 0.069$ ).

## Questionnaire Results

The proportion of readers satisfied or very satisfied with the LLM-assisted diagnostic workflow increased from 55.6% (5/9) to 88.9% (8/9) following the tutorial. 55.6% (5/9) of readers indicated they would consider using the tool in clinical practice before the training, compared to 88.9% (8/9) after the training. 88.9% (8/9) of readers found the tutorial helpful or very helpful.

## Reader Feedback and Observations

Before the tutorial, most readers tended to use keyword-based queries rather than detailed instructions, similar to how conventional search engines operate. In general, the LLM tool was found particularly useful in generating an initial list of possible differential diagnoses to be evaluated through additional searches. In addition, the possibility to pose follow-up questions to initial query results was perceived as an advantage over conventional search engines. When using the provided prompt template, readers overwhelmingly appreciated the concise tabular format of the results which also included the rationale for the suggestion. However, some readers struggled to formulate accurate image descriptions, owing to their insufficient knowledge of neuroanatomy and brain MRI sequences.

## LLM Response Evaluation

A total of 413 LLM queries in 169 patient cases were examined (2.44 queries per case). In 11 out of 180 cases, LLM logs were not available because the user did not perform any LLM queries or because logs could not be retrieved due to technical errors of PerplexityAI. 5.8% of queries included incorrect inputs, such as inaccurate descriptions of finding locations or imaging characteristics. 7.3% of queries included screenshots of MRI findings.

The proportion of instruction-based queries increased substantially from 35.2% to 95.4% after the tutorial, while keyword queries decreased inversely from 64.8% to 4.6%. Whereas the majority of queries prior to the training were classified as “Other” (58.4%), most queries following the tutorial were directed at relevant differential diagnoses (56.7%) and sample images of those (34.0%). Overall, 45.3% of sources indicated by

PerplexityAI were peer-reviewed journal articles, with only minor differences between queries before and after the tutorial (49.2% and 40.6% each).

Hallucinations were observed in 5.1% of LLM queries (12.4% of cases; 20 responses in total). The LLM tutorial resulted in only a minor change in hallucination frequency (Pre-Training: 4.6%, Post-Training: 5.7%). 35.0% of hallucinations involved the misinterpretation of MRI screenshots provided as input or the LLM returning sample MRI images irrelevant to the context of the query. Hallucination details are provided in Supplement 2. Misinterpretations were found in 1.0% of queries (2.4% of cases), while clarifications occurred in 1.2% of queries (3.0% of cases).

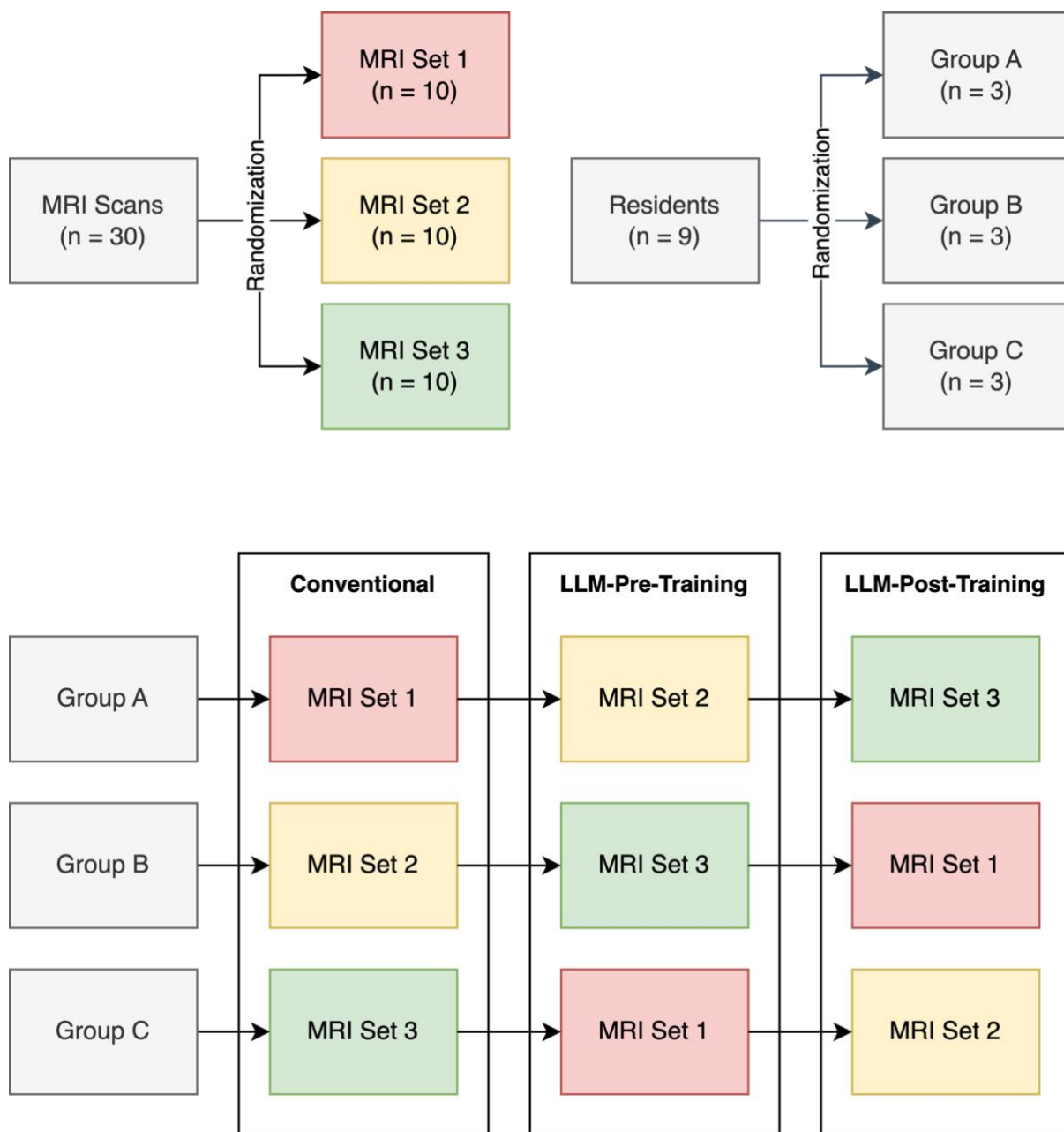


Figure 1: Study Design.

You are a senior neuroradiologist. Below, you will find information regarding a brain MRI scan. Based on this information, identify the three most likely differential diagnoses, ranked by their likelihood. Present your findings in a table format with the following columns: "Rank", "Differential Diagnosis", and "Explanation".

Female patient in her early 30s. Headache, history of Cesarean section and HELLP syndrome 3 weeks ago, HSV pneumonia. Well-defined, round T2-hyperintense lesion in the splenium of the corpus callosum with distinct diffusion restriction.

#### Sources

Splenia Lesions of the Corpus Callosum...  
ncbi.nlm · 1

Reversible lesion in the splenium of the corpus...  
ncbi.nlm · 2

Postpartum headache - diagnosis and treatmen...  
ncbi.nlm · 3

View 2 more

#### Answer

Based on the provided clinical and MRI findings, the differential diagnoses for a well-defined, round T2-hyperintense lesion in the splenium of the corpus callosum with distinct diffusion restriction in a female patient in her early 30s with a history of Cesarean section, HELLP syndrome, and HSV pneumonia are as follows:

| Rank | Differential Diagnosis                              | Explanation   |
|------|---|---|
| 1    | Cytotoxic Lesions of the Corpus Callosum (CLOCCs)   | Given the distinct diffusion restriction and the T2-hyperintense nature of the lesion, CLOCCs are a primary consideration. These lesions are associated with various conditions, including infections and metabolic disturbances, which could be related to the patient's recent medical history <sup>4</sup> . |
| 2    | Posterior Reversible Encephalopathy Syndrome (PRES) | Although PRES typically presents with more extensive edema affecting the parieto-occipital regions, isolated splenial involvement can occur. The patient's history of HELLP syndrome, a condition associated with eclampsia, increases the likelihood of PRES <sup>3</sup> .                                    |
| 3    | Reversible Splenial Lesion Syndrome (RESLES)        | This condition, often triggered by infections or abrupt changes in homeostasis, presents with similar MRI findings. The patient's recent history of HSV pneumonia suggests a possible infectious trigger for RESLES <sup>2</sup> .  |

These differential diagnoses are ranked based on the specific MRI characteristics of the lesion, the patient's recent medical history, and the commonality of these conditions in similar clinical scenarios.

Share Rewrite

GPT-4 TURBO

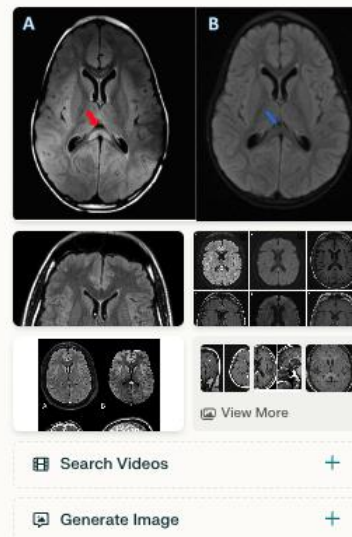
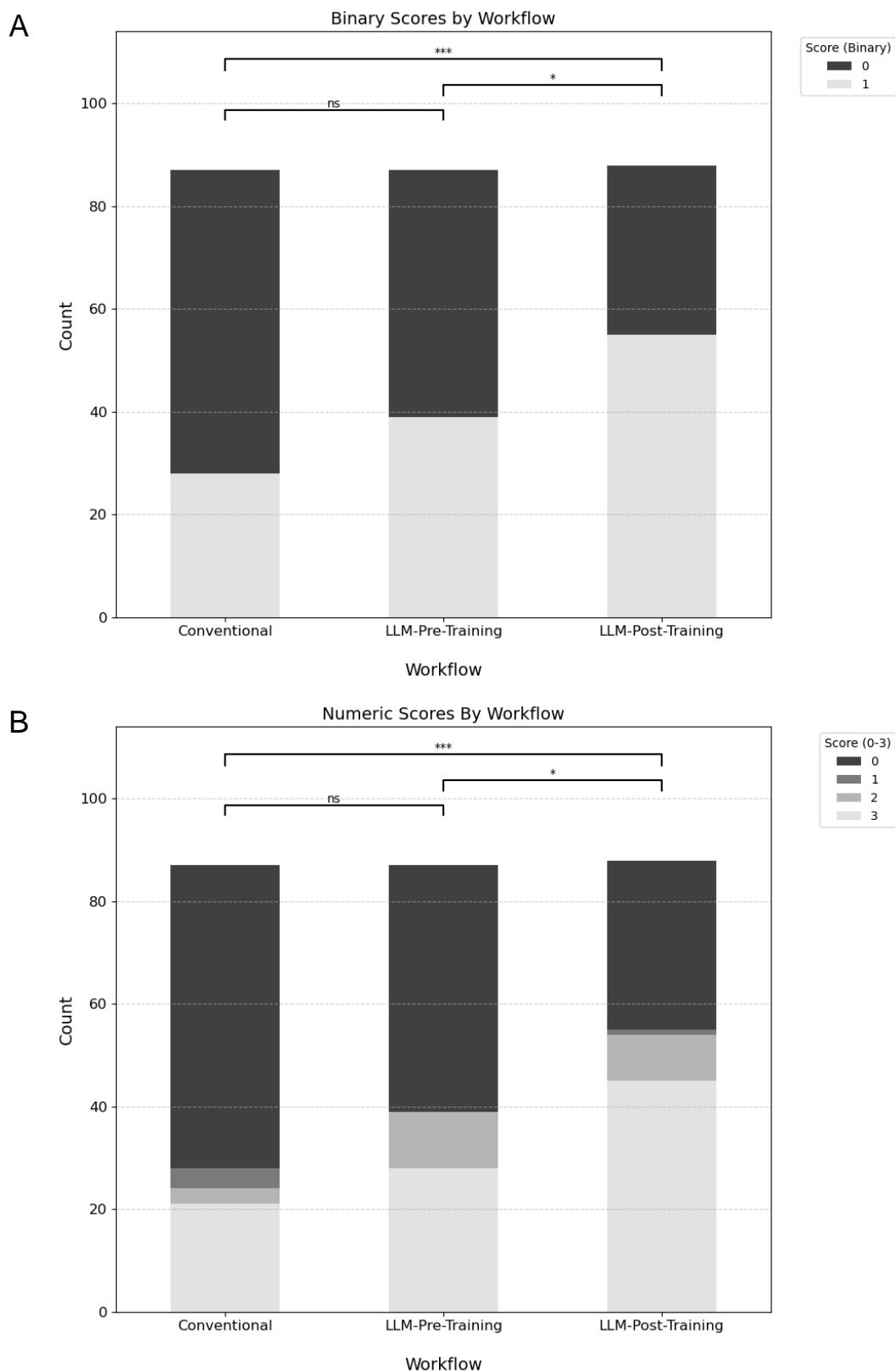


Figure 2: Screenshot of the PerplexityAI user interface. The correct diagnosis in this case was cytotoxic lesion of the corpus callosum (CLOCC).





**Figure 3: Diagnostic performance by workflow.** *A: Binary scores. Responses were classified as either correct (1) or incorrect (0). B: Numeric scores. Responses were assigned a score between 0 and 3, depending on the rank of the correct diagnosis within the response (3: correct diagnosis ranked first, 0: correct diagnosis not included in response). \*  $p < 0.05$ . \*\*\*  $p < 0.001$ . ns: not significant.*

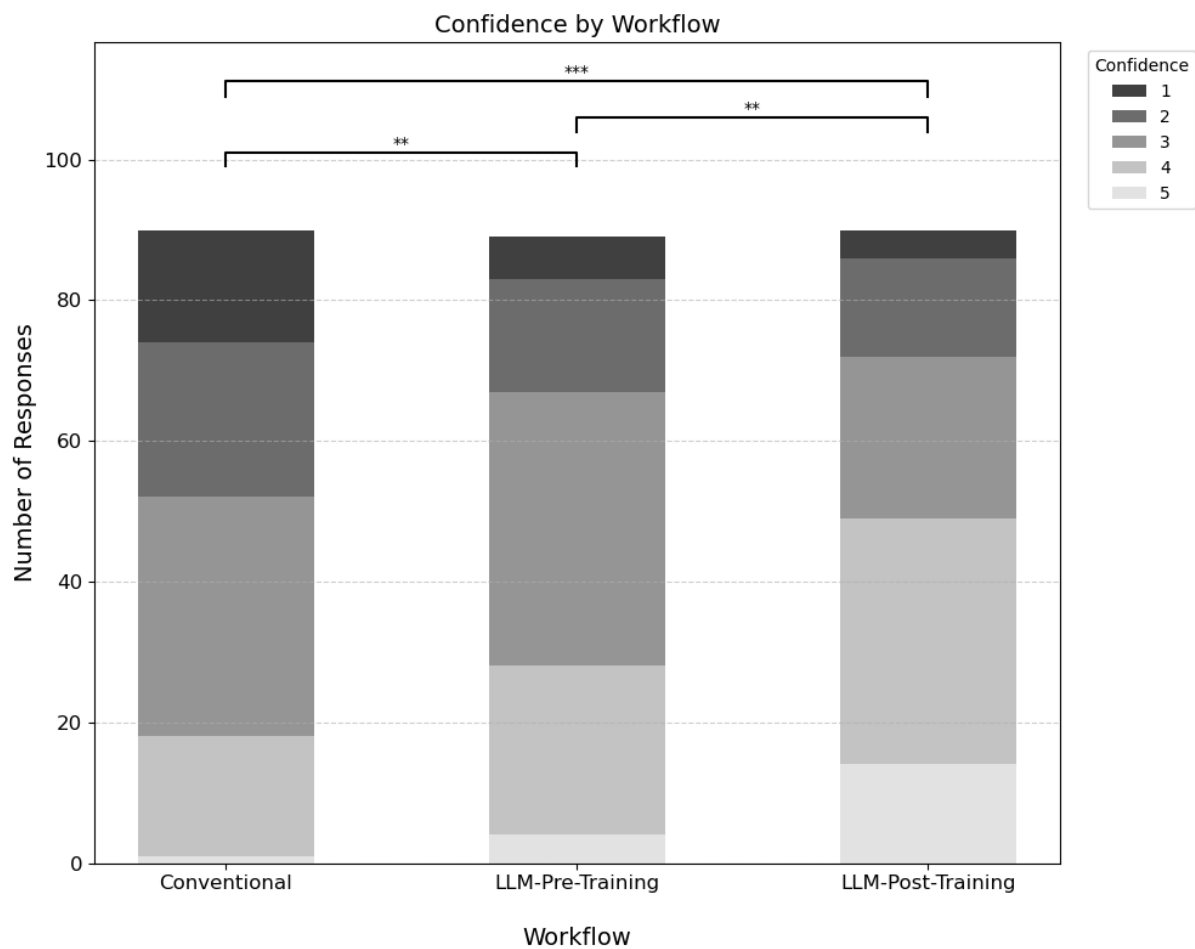


Figure 4: Confidence (5-point Likert scale rating) by workflow. \*\*  $p < 0.01$ . \*\*\*  $p < 0.001$ .

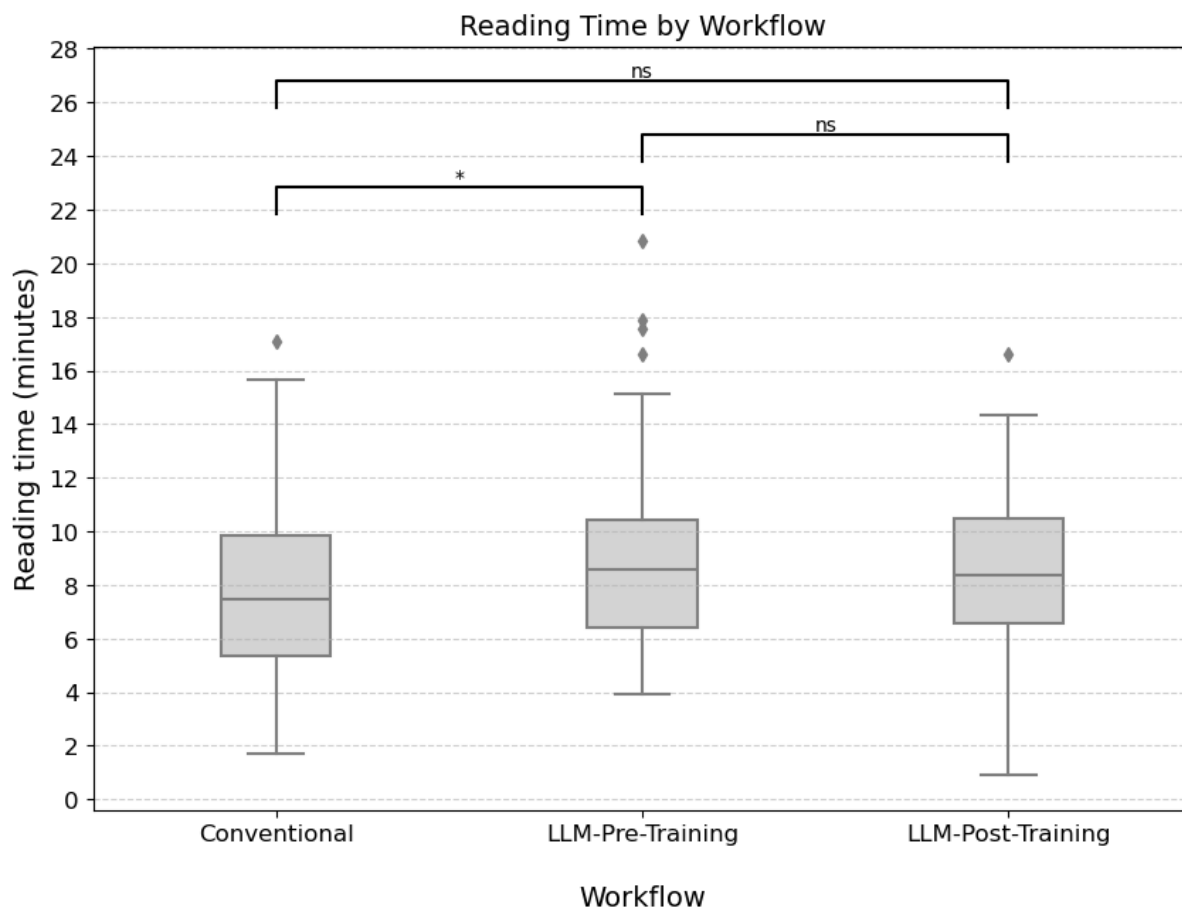
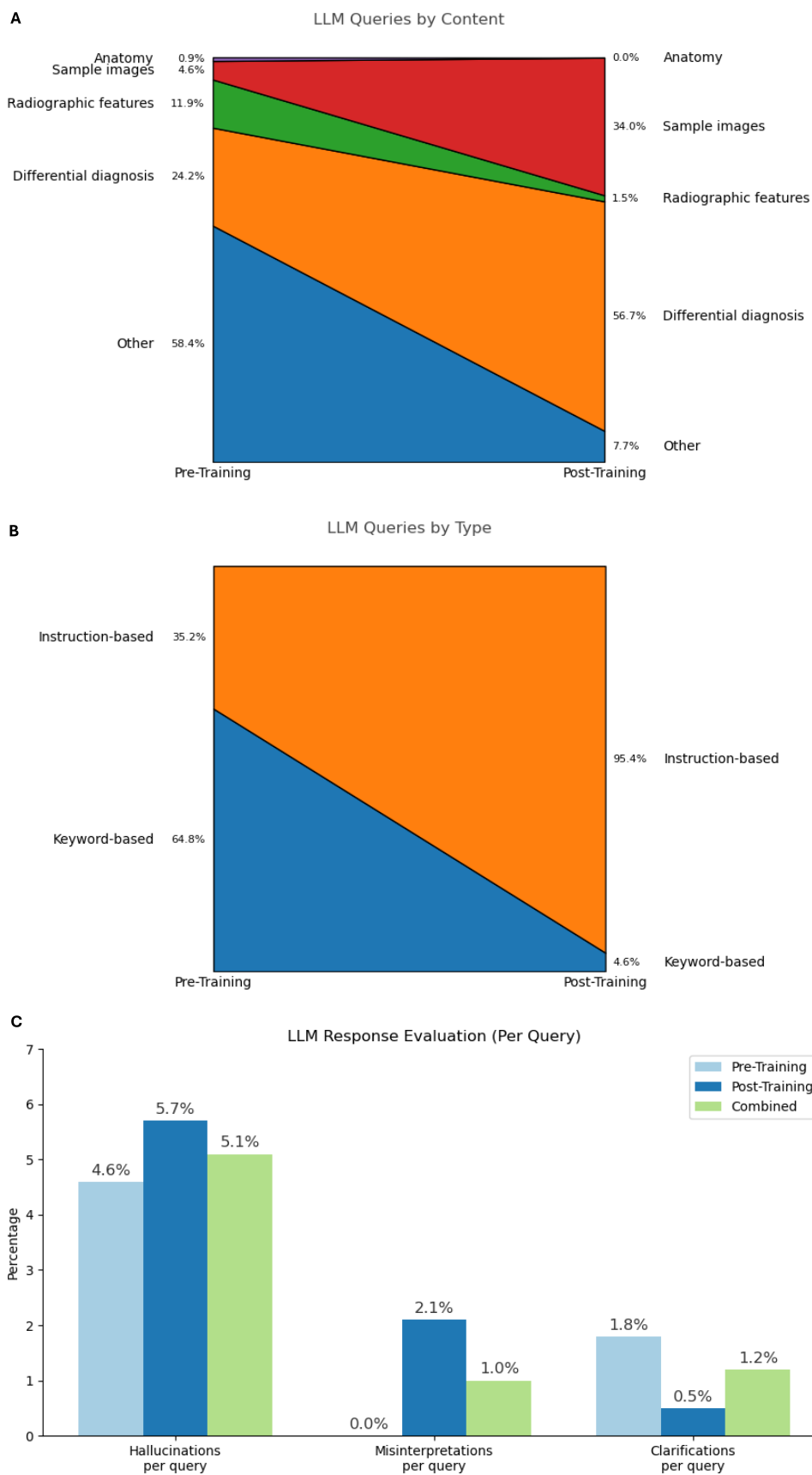


Figure 5: Reading time by workflow. \*  $p < 0.05$ . ns: not significant.



**Figure 6: Evaluation of LLM queries and responses. A: LLM queries by content. B: LLM queries by type. C: Relative frequency of hallucinations, misinterpretations and clarifications per query.**

| Trait   | Value                             |
|---|-----------------------------------|
| Total Number of Readers                         | 9                                 |
| Gender Distribution                             | 7 Male (77.8%) / 2 Female (22.2%) |
| Mean radiology experience (in months)           | 18.56 ± 15.93                     |
| Mean neuroradiology experience (in months)      | 2.22 ± 2.05                       |
| Readers who have used LLMs before               | 6/9 (66.7%)                       |
| Readers who have used LLMs for diagnosis before | 1/9 (11.1%)                       |

*Table 1: Overview of readers.*

| Comparison                            | Chi2 Statistic | p       | p (adjusted)      |
|---------------------------------------|----------------|---------|-------------------|
| LLM-Pre-Training vs LLM-Post-Training | 4.81           | 0.028   | <b>0.042</b>      |
| LLM-Pre-Training vs Conventional      | 2.43           | 0.119   | 0.119             |
| LLM-Post-Training vs Conventional     | 14.93          | < 0.001 | <b>&lt; 0.001</b> |

*Table 2: Pairwise testing for inter-group differences in binary scores. Adjusted p-values have been corrected for a false-discovery rate of 0.05.*

| Comparison                            | U Statistic | p       | p (adjusted)      |
|---------------------------------------|-------------|---------|-------------------|
| LLM-Pre-Training vs LLM-Post-Training | 3049.5      | 0.011   | <b>0.016</b>      |
| LLM-Pre-Training vs Conventional      | 4267.5      | 0.092   | 0.092             |
| LLM-Post-Training vs Conventional     | 5057.5      | < 0.001 | <b>&lt; 0.001</b> |

*Table 3: Pairwise testing for inter-group differences in numeric scores. Adjusted p-values have been corrected for a false-discovery rate of 0.05.*

## Discussion

This study evaluated the impact of a structured 10-minute LLM tutorial on the performance of radiology residents in LLM-assisted brain MRI differential diagnosis. We found that readers displayed higher performance, confidence levels and overall satisfaction after completing the tutorial. Compared to differential diagnosis supported by conventional internet search, both LLM-assisted workflows resulted in better performance, although only the post-training workflow showed a statistically significant difference.

Analysis of reader-LLM interactions revealed that following the tutorial, almost all queries were phrased as specific instructions, whereas most queries before the training consisted of mere keywords, resembling conventional search engine queries. This observation is consistent with “Jakob’s Law” which is a well-known phenomenon in user experience (UX) stating that users prefer systems to behave like other familiar ones (26). Similar to hallucination rates described previously (25), we found statements inconsistent with widely accepted medical knowledge in 5.1% of LLM responses. Many of these involved incorrect interpretations of MRI screenshots provided as input, confirming earlier studies demonstrating low performance of current state-of-the-art LLMs in diagnostic tasks based on radiological images (27–30). Interestingly, hallucinations were even found with PerplexityAI, which – unlike other Chatbots such as ChatGPT - combines LLMs with real-time information retrieval from the internet to support its answers with relevant sources (31). Additional research is needed to develop further safeguarding measures against potentially harmful effects of hallucinations in clinical LLM applications. Notably, in 5.8% of queries misleading responses were generated not because of undesired LLM behavior, but due to

incorrect finding descriptions provided by readers, emphasizing the essential role of conventional radiological skills in effectively employing LLMs for diagnostic tasks.

Our findings indicate that even minor, low-cost educational interventions for LLMs can yield remarkable outcomes, and support the notion that courses focused on the practical application of AI should become a core part of medical curricula and training programs (32–34). Yet, given the novelty of the technology, validated educational content on the effective utilization of LLMs for specific clinical tasks is extremely scarce. The tutorial provided to the readers in this work was based on two previous studies specifically focusing on human-LLM collaboration and prompt engineering in brain MRI differential diagnosis (23,24). As the corpus of scientific evidence on LLM applications grows, medical societies should provide guidelines and courses on their appropriate use. Furthermore, platforms should be created to allow for healthcare professionals to exchange validated and effective prompts, similar to previous initiatives by radiological societies for sharing structured reporting templates (35–37).

Unlike diagnostic performance, reading times did not improve with either of the LLM-assisted workflows. As prior work on AI-based image analysis algorithms and structured reporting has illustrated (38,39), integration of technologies into the local IT infrastructure is critical for user acceptance and can boost efficiency. Vendors of radiology reporting solutions and Picture Archiving and Communication Systems (PACS) should explore ways to seamlessly embed LLM-based features supporting differential diagnosis and other tasks.

## Limitations

The following limitations need to be acknowledged.

First, only radiology residents with very little neuroradiology experience were included. This study design ensured that most cases could not be solved by readers with prior knowledge alone and allowed to investigate the isolated effect of distinct web research workflows, but limits generalizability of the findings to more experienced readers. Yet, our observation that inexperienced readers showed performance improvements ensuing the LLM tutorial despite struggling to formulate image descriptions suggests that moderately experienced readers, who are proficient enough to create accurate finding descriptions but not yet skilled enough to conduct differential diagnoses without assistance, might benefit even more.

Second, the findings in question were presented with annotations to isolate readers' classification performance, but this approach reduced the realism of the scenario. It is likely that in actual practice, some of the more subtle findings would have been missed, as remarked by several readers.

Third, this study employed only a single LLM (GPT-4 by OpenAI) accessed through a specialized search engine (PerplexityAI). Future studies should compare several closed-source and open-source LLMs with respect to their utility in supporting radiology readers in differential diagnosis, including ones fine-tuned with domain-specific training data.

In conclusion, a concise but structured 10-minute LLM tutorial increased performance and confidence levels in LLM-assisted brain MRI differential diagnosis among radiology residents. These findings highlight the considerable benefits that even low-cost, low-effort educational interventions on LLMs can provide.



## References

1. Bhayana R. Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. *Radiology*. Radiological Society of North America Inc.; 2024;310(1). doi: 10.1148/RADIOL.232756/ASSET/IMAGES/LARGE/RADIOL.232756.FIG6.JPG.
2. Gertz RJ, Bunck AC, Lennartz S, et al. GPT-4 for Automated Determination of Radiologic Study and Protocol Based on Radiology Request Forms: A Feasibility Study. *Radiology*. Radiological Society of North America Inc.; 2023;307(5). doi: 10.1148/RADIOL.230877.
3. Arachchige ASPM. Can We Use Large Language Models to Guide the Use of Contrast Media in Radiology? Reply to Kaba et al. *Acad Radiol*. Elsevier; 2024;0(0). doi: 10.1016/J.ACRA.2023.12.044.
4. Rau A, Rau S, Zöller D, et al. A Context-based Chatbot Surpasses Radiologists and Generic ChatGPT in Following the ACR Appropriateness Guidelines. *Radiology*. Radiological Society of North America Inc.; 2023;308(1). doi: 10.1148/RADIOL.230970.
5. Kumar Sarangi P, Irodi A, Panda S, Swapnesh Kumar Nayak D, Mondal H. Radiological Differential Diagnoses Based on Cardiovascular and Thoracic Imaging Patterns: Perspectives of Four Large Language Models. *Indian Journal of Radiology and Imaging*. 2023; doi: 10.1055/s-0043-1777289.
6. Horiuchi D, Tatekawa H, Shimono T, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. *Neuroradiology*. *Neuroradiology*; 2024;66(1). doi: 10.1007/S00234-023-03252-4.
7. Ueda D, Mitsuyama Y, Takita H, et al. ChatGPT's Diagnostic Performance from Patient History and Imaging Findings on the Diagnosis Please Quizzes. *Radiology*. Radiological Society of North America Inc.; 2023;308(1). doi: 10.1148/RADIOL.231040.
8. Kottlors J, Bratke G, Rauen P, et al. Feasibility of Differential Diagnosis Based on Imaging Patterns Using a Large Language Model. *Radiology*. Radiological Society of North America Inc.; 2023;308(1). doi: 10.1148/radiol.231167.

9. Bhayana R, Bleakney RR, Krishna S. GPT-4 in Radiology: Improvements in Advanced Reasoning. *Radiology*. Radiological Society of North America Inc.; 2023;307(5). doi: 10.1148/RADIOL.230987.
10. Wu J, Kim Y, Keller EC, et al. Exploring Multimodal Large Language Models for Radiology Report Error-checking. 2023; <https://arxiv.org/abs/2312.13103v1>. Accessed January 14, 2024.
11. Ziegelmeier S, Marka AW, Lenhart N, et al. Evaluation of GPT-4's Chest X-Ray Impression Generation: A Reader Study on Performance and Perception. *J Med Internet Res*. JMIR Publications Inc.; 2023;25(1):e50865. doi: 10.2196/50865.
12. Sun Z, Ong H, Kennedy P, et al. Evaluating GPT-4 on Impressions Generation in Radiology Reports. *Radiology*. Radiological Society of North America Inc.; 2023;307(5). doi: 10.1148/RADIOL.231259.
13. Huang J, Yang DM, Rong R, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. *npj Digital Medicine* 2024 7:1. Nature Publishing Group; 2024;7(1):1–13. doi: 10.1038/s41746-024-01079-8.
14. 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. Anzai Y, editor. *Radiology*. Radiological Society of North America ; 2024;311(1). doi: 10.1148/RADIOL.232741.
15. Bhayana R, Elias G, Datta D, Bhabra N, Deng Y, Krishna S. Use of GPT-4 With Single-Shot Learning to Identify Incidental Findings in Radiology Reports. *AJR Am J Roentgenol*. American Roentgen Ray Society ; 2024; doi: 10.2214/AJR.23.30651/SUPPL\_FILE/23\_30651\_SUPPL.PDF.
16. Wu J, Kim Y, Wu H. Hallucination Benchmark in Medical Visual Question Answering. 2024; <https://www.openai.com/gpt-4>. Accessed January 26, 2024.
17. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. *Crit Care*. BioMed Central Ltd; 2023;27(1):1–2. doi: 10.1186/S13054-023-04393-X.
18. Shen Y, Heacock L, Elias J, et al. ChatGPT and Other Large Language Models Are Double-edged Swords. *Radiology*. Radiological Society of North America Inc.; 2023;307(2). doi: 10.1148/RADIOL.230163.

19. Bhayana R, Biswas S, Cook TS, et al. From Bench to Bedside With Large Language Models: AJR Expert Panel Narrative Review. *AJR Am J Roentgenol.* *AJR Am J Roentgenol*; 2024; doi: 10.2214/AJR.24.30928.
20. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *J Med Syst.* Springer; 2023;47(1):1–5. doi: 10.1007/S10916-023-01925-4/TABLES/2.
21. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large Language Models in Medicine: The Potentials and Pitfalls : A Narrative Review. *Ann Intern Med.* *Ann Intern Med*; 2024;177(2):210–220. doi: 10.7326/M23-2772.
22. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. *Communications Medicine* 2023 3:1. Nature Publishing Group; 2023;3(1):1–8. doi: 10.1038/s43856-023-00370-1.
23. Kim SH, Schramm S, Berberich C, et al. Human-AI Collaboration in Large Language Model-Assisted Brain MRI Differential Diagnosis: A Usability Study. medRxiv. Cold Spring Harbor Laboratory Press; 2024;2024.02.05.24302099. doi: 10.1101/2024.02.05.24302099.
24. Schramm S, Preis S, Metz M-C, et al. Impact of Multimodal Prompt Elements on Diagnostic Performance of GPT-4(V) in Challenging Brain MRI Cases. medRxiv. Cold Spring Harbor Laboratory Press; 2024;2024.03.05.24303767. doi: 10.1101/2024.03.05.24303767.
25. Siepmann R, Huppertz M, Rastkhiz A, et al. The virtual reference radiologist: comprehensive AI assistance for clinical image reading and interpretation. *Eur Radiol.* Springer Science and Business Media Deutschland GmbH; 2024;1–15. doi: 10.1007/S00330-024-10727-2/TABLES/3.
26. Yablonski J. *Laws of UX.* “ O’Reilly Media, Inc.”; 2024.
27. Wu C, Lei J, Zheng Q, et al. Can GPT-4V(ision) Serve Medical Applications? Case Studies on GPT-4V for Multimodal Medical Diagnosis. 2023; <https://arxiv.org/abs/2310.09909v3>. Accessed February 22, 2024.
28. Horiuchi D, Tatekawa H, Oura T, et al. Comparison of the diagnostic accuracy among GPT-4 based ChatGPT, GPT-4V based ChatGPT, and radiologists in musculoskeletal radiology. medRxiv. Cold Spring Harbor Laboratory Press; 2023;2023.12.07.23299707. doi: 10.1101/2023.12.07.23299707.

29. Brin D, Sorin V, Barash Y, et al. Assessing GPT-4 Multimodal Performance in Radiological Image Analysis. medRxiv. Cold Spring Harbor Laboratory Press; 2023;2023.11.15.23298583. doi: 10.1101/2023.11.15.23298583.
30. Deng J, Heybati K, Shammass-Toma M. When vision meets reality: Exploring the clinical applicability of GPT-4 with vision. Clin Imaging. Elsevier; 2024;108:110101. doi: 10.1016/J.CLINIMAG.2024.110101.
31. Perplexity Accelerates Foundation Model Training by 40% with Amazon SageMaker HyperPod | Perplexity Case Study | AWS. . <https://aws.amazon.com/solutions/case-studies/perplexity-case-study/>. Accessed July 1, 2024.
32. Paranjape K, Schinkel M, Panday RN, Car J, Nanayakkara P. Introducing Artificial Intelligence Training in Medical Education. JMIR Med Educ. JMIR Publications Inc.; 2019;5(2). doi: 10.2196/16048.
33. Ngo B, Nguyen D, Vansonnenberg E. The Cases for and against Artificial Intelligence in the Medical School Curriculum. Radiol Artif Intell. Radiological Society of North America; 2022;4(5). doi: 10.1148/RYAI.220074.
34. Li Q, Qin Y. AI in medical education: medical student perception, curriculum recommendations and design suggestions. BMC Med Educ. BioMed Central Ltd; 2023;23(1):1–8. doi: 10.1186/S12909-023-04700-8/FIGURES/2.
35. Pinto Dos Santos D, Hempel JM, Mildenerger P, Klöckner R, Persigehl T. Structured Reporting in Clinical Routine. RoFo Fortschritte auf dem Gebiet der Röntgenstrahlen und der Bildgebenden Verfahren. Georg Thieme Verlag; 2019;191(1):33–39. doi: 10.1055/a-0636-3851.
36. Kohli M, Alkasab T, Wang K, et al. Bending the Artificial Intelligence Curve for Radiology: Informatics Tools From ACR and RSNA. Journal of the American College of Radiology. Elsevier; 2019;16(10):1464–1470. doi: 10.1016/J.JACR.2019.06.009.
37. ESR. ESR paper on structured reporting in radiology—update 2023. Insights Imaging. Springer; 2023;14(1). doi: 10.1186/S13244-023-01560-0.
38. Kim SH, Mir-Bashiri S, Matthies P, Sommer W, Nörenberg D. Integration of structured reporting into the routine radiological workflow. Radiologe. Springer Medizin; 2021;61(11):1005–1013. doi: 10.1007/S00117-021-00917-0/FIGURES/8.

39. Dikici E, Bigelow M, Prevedello LM, White RD, Erdal BS. Integrating AI into radiology workflow: levels of research, production, and feedback maturity. *Journal of Medical Imaging. Society of Photo-Optical Instrumentation Engineers*; 2020;7(1):1. doi: 10.1117/1.JMI.7.1.016502.

| ID   | Age Range | Sex | Correct Diagnosis                                   | Condensed Medical History  |
|------|-----------|-----|---|--|
| 1.1  | 36 - 40   | F   | Midline glioma                                      | Intracranial mass in external CT scan, headache  |
| 1.2  | 31 - 35   | F   | ICA dissection                                      | Recurring visual impairment on the left side, headache   |
| 1.3  | 31 - 35   | M   | Hemangioblastoma                                    | Dizziness, headache  |
| 1.4  | 41 - 45   | M   | Chiari malformation type 1                          | Right-sided headache, dizziness, right facial hypaesthesia   |
| 1.5  | 41 - 45   | M   | Periorbital phlegmone                               | Periorbital swelling, headache, fever  |
| 1.6  | 56 - 60   | F   | Cerebral venous thrombosis                          | Acute desorientation   |
| 1.7  | 51 - 55   | F   | Chondrosarcoma                                      | Newly diagnosed left sixth nerve palsy with diplopia, severe headaches for 2 days, left facial numbness for several weeks  |
| 1.8  | 36 - 40   | F   | Tumefactive multiple sclerosis (MS)                 | Sensory-motor hemiparesis left   |
| 1.9  | 46 - 50   | M   | Multinodular and vacuolating neuronal tumor (MVNT)  | Subacute ptosis, right oculomotor palsy, diplopia  |
| 1.10 | 36 - 40   | M   | Cerebral cavernous venous malformation              | Incidental finding in left temporal lobe in external imaging   |
| 2.1  | 31 - 35   | F   | Cytotoxic lesions of the corpus callosum (CLOCC)    | Headache, history of Cesarean section and HELLP syndrome 3 weeks ago, HSV pneumonia  |
| 2.2  | 61 - 65   | M   | Cerebral amyloid angiopathy                         | Aphasia since yesterday, mild right hemiparesis  |
| 2.3  | 41 - 45   | F   | Echinococcosis                                      | Headache, vomiting, mood disturbances  |
| 2.4  | 41 - 45   | M   | Pontine myelinolysis                                | Suspected brain abscess, patent foramen ovale  |
| 2.5  | 36 - 40   | M   | Creutzfeldt Jakob disease                           | History of alcoholic liver cirrhosis, suspicion of hyperactive delirium  |
| 2.6  | 41 - 45   | M   | Ecchordosis physaliphora                            | Gait ataxia, dysarthria, depression  |
| 2.7  | 51 - 55   | M   | Miller-Fisher syndrome                              | Incidental clival finding in external imaging, patient is asymptomatic   |
| 2.8  | 61 - 65   | F   | Progressive multifocal encephalopathy (PML)         | Speech impairment since two weeks, double vision since three days, progressive swallowing disorder                         |
| 2.9  | 56 - 60   | M   | Limbic encephalitis                                 | Progressive right-sided hemiparesis, history of chronic lymphocytic leukemia   |
| 2.10 | 41 - 45   | F   | Spinocerebellar ataxia                              | Organic psychosyndrome for 3 months, psychomotor retardation, dysarthria   |
| 3.1  | 61 - 65   | M   | Superficial siderosis                               | Progressive gait ataxia, upper limb ataxia for 2 years   |
| 3.2  | 21 - 25   | F   | Moya Moya disease                                   | Severe left sensorineural hearing loss   |
| 3.3  | 31 - 35   | M   | Sarcoma   | Left hemiparesis for several weeks   |
| 3.4  | 56 - 60   | F   | Huntington's disease                                | Scalp lesion Ddx lipoma Ddx atheroma, Pt. noticed swelling a few months ago, has grown since                               |
| 3.5  | 56 - 60   | M   | Subependymoma                                       | Repeated orofacial dyskinesias, parathymic behavior. Generalized brain atrophy in CT.                                      |
| 3.6  | 36 - 40   | M   | Toxoplasmosis                                       | Ddx neurodegenerative disease  |
| 3.7  | 26 - 30   | M   | Central neurocytoma                                 | Incidental intracranial mass lesion in external images   |
| 3.8  | 21 - 25   | F   | Benign enhancing foramen magnum lesion              | Patient is asymptomatic  |
| 3.9  | 56 - 60   | F   | Posterior reversible encephalopathy syndrome (PRES) | Dysarthria, dysphagia, psychomotor slowing, suspected HIV infection, intracerebral lesions                                 |
| 3.10 | 76 - 80   | M   | Diffuse axonal injury                               | Headache and papilledema   |
|      |           |     |   | Ski injury with head impact, retrograde amnesia, initial confusion, subsequent improvement, MRI preferred due to young age |
|      |           |     |   | Confusion, immunosuppression post-kidney transplant  |
|      |           |     |   | Recent history of high-speed trauma with traumatic SAH and hemotopneumothorax  |

*Supplement 1: Case Overview.*

| Reader Prompt   | Hallucinatory Output  | Explanation  |
|---|---|--|
| <p>[Standard Prompt Introduction]<br/> Male in his 30s, scalp lesion differential lipoma differential atheroma, patient noticed swelling a few months ago and the lesion has grown since</p> <p>The brain MRI shows a large extradural lesion emerging from the dura of the left frontal lobe. It measures roughly 33 x 48 mm on the coronal view. The lesion expands out of the calotte, destroying the bone, and extends out of the skull. The skin is intact. The contrast image shows inhomogeneous gadolinium enhancement of the lesion. The lesion seems composed of cystic components. There is compressive effect upon the parenchyma of the left hemisphere. The midline is shifted slightly to the right.</p> | <p>"The inhomogeneous enhancement with gadolinium and the presence of cystic components are also characteristic features of meningiomas"</p>  | <p>Meningiomas typically exhibit intense and homogeneous contrast enhancement. The LLM response indicates that inhomogeneous enhancement and cystic components are characteristic features of meningioma, which is incorrect.</p>        |
| <p>[Standard Prompt Introduction]<br/> Female in her 20s, ski injury with head impact, retrograde amnesia, initial confusion, subsequent improvement</p> <p>The brain MRI shows a small roundly-shaped FLAIR hyperintense solid mass infratentorially anterior to the left cerebellar hemisphere. The contours are patchy. It is surrounded by cerebrospinal fluid. There is no diffusion restriction or microhemorrhage. The brain nerves are emerging from the brain stem in close proximity to the lesion but are not compressed.</p>  | <p>"[...] The description of the lesion being surrounded by cerebrospinal fluid and its infratentorial location anterior to the cerebellum could suggest an arachnoid cyst."</p>  | <p>Arachnoid cysts are not FLAIR-hyperintense (as indicated in the prompt).</p>  |
| <p>[Standard Prompt Introduction]<br/> female in her 50s, confusion, immunosuppression following kidney transplant. Bilateral occipital FLAIR hyperintensities that do not involve the cortex. TOF MRA without any pathological finding. No diffusion restriction</p>   | <p>"The symmetric FLAIR hyperintensities in the occipital lobes and cerebellum are typical of PRES, although the absence of diffusion restriction is less typical"</p>  | <p>DWI signal is usually normal in PRES, restricted diffusion is presented in a quarter of cases (Radiopaedia).</p>  |
| <p>Generate a list of differential diagnoses for a MRI scan of a patient with the following incidental imaging finding on brain MRI: [images of subependymoma in the left lateral ventricle]</p>  | <p>"Based on the provided MRI images, which appear to show an incidental finding within the brain, a list of differential diagnoses could include:<br/> [...]<br/> 3. Pineal cyst: A benign, fluid-filled sac located in the pineal gland in the brain.<br/> 4. Rathke's cleft cyst: A noncancerous cyst that occurs in the pituitary gland."</p>   | <p>Based on two images showing an intraventricular lesion, the LLM mentioned "pineal cyst" and "Rathke's cleft cyst" as differential diagnoses, although the respective regions were not shown in the images.</p>                        |
| <p>Generate a list of differential diagnoses for a MRI scan of a patient with dysarthria, dysphagia and HIV and the following brain lesion:</p>   | <p>"Progressive Multifocal Leukoencephalopathy (PML): Caused by the JC virus, PML can occur in people with severe immunodeficiency, including those with HIV/AIDS. MRI findings often show non-enhancing white matter lesions without mass effect."</p>   | <p>The provided screenshot did not show any white matter lesions.</p>  |
| <p>Generate a list of differential diagnoses for a MRI scan of a patient in his 30s with a headache and the following imaging finding: [image]</p>  | <p>"Based on the provided MRI scan and the symptoms of a headache, here is a list of potential differential diagnoses:<br/> [...]<br/> Cerebral venous sinus thrombosis: A blood clot in the brain's venous sinuses, which can cause increased intracranial pressure and headaches.<br/> Encephalitis: Inflammation of the brain, often caused by an infection.<br/> Multiple sclerosis: A disease of the central nervous system that can present with lesions in the brain and cause headaches."</p> | <p>In a case of an intraventricular lesion (central neurocytoma), several differential diagnoses not applicable to the intraventricular location are mentioned (cerebral venous sinus thrombosis, encephalitis, multiple sclerosis).</p> |

|  |   |  |
|--|---|--|
| <p>Generate a list of differential diagnoses for a MRI scan of a young female patient after skiing accident with the following imaging finding: [Image]</p> <p>show an MRI example of Guillain-Barré Syndrome (GBS) where the facial nerve is enhanced</p> | <p>"[...]<br/>Ligament Injury: Injuries to the cruciate or collateral ligaments are common in skiing accidents due to twisting forces on the knee.<br/>Meniscal Tear: The menisci are cartilage structures in the knee that can be torn during accidents involving forceful twisting or rotation of the knee.<br/>Bone Contusion or Fracture: Direct impact or stress from an awkward fall can lead to bone bruises or fractures.<br/>Muscle Strain or Tear: The force exerted on muscles during a skiing accident can cause strains or tears, particularly in the quadriceps or hamstrings.<br/>Tendon Injury: Tendons, such as the patellar tendon, may be strained or torn if overstretched during an accident.<br/>[...]"</p>   | <p>In a case of a foramen magnum lesion, multiple differential diagnoses not applicable to the location of the lesion are suggested (meniscal tear, bone contusion, muscle strain, patellar tendon injury, bursitis).<br/>Images of spinal GBS manifestations are provided.</p>  |
| <p>Bilateral diffusion disorder with FLAIR correlate in the dorsal corpus callosum</p>   | <p>"Such MRI findings can occur in a variety of diseases and conditions, including<br/>Mild Encephalopathy with Reversible Splenial Lesion (MERS): a form of mild encephalopathy characterized by reversible lesions in the splenium of the corpus callosum. This disease is often associated with infections and can occur in children and adults.<br/>Cytotoxic Lesions of the Corpus Callosum (CLOCCs): This category includes various conditions that lead to cytotoxic edema in the corpus callosum, often associated with infections, metabolic disorders, or abrupt discontinuation of antiepileptic drugs.<br/>[...]"</p>   | <p>Lists "Mild Encephalopathy with Reversible Splenial Lesion (MERS)" and "Cytotoxic Lesions of the Corpus Callosum (CLOCCs)" as two distinct differential diagnoses, although these are synonyms.</p>   |
| <p>What are the reasons for diffusion disorders in the pons without contrast agent uptake?</p>   | <p>"Mild Encephalopathy with Reversible Splenial Lesion (MERS): MERS is a rare disease characterized by reversible lesions in the corpus callosum, but can also affect other brain regions including the pons. These lesions show restricted diffusion without contrast uptake, indicating reversible intracellular swelling."<br/>"Movement disorders: The basal ganglia play a central role in the regulation of movement. Disorders such as restless legs syndrome, periodic movement disorders of the extremities and stereotypies can be associated with changes in the basal ganglia."</p>  | <p>MERS / CLOCC is by definition located in the corpus callosum. The LLM response inaccurately indicates that it can be located in the pons.</p>   |
| <p>which causes can lead to a bilateral diffusion disorder in the basal ganglia</p>  | <p>"Movement disorders: The basal ganglia play a central role in the regulation of movement. Disorders such as restless legs syndrome, periodic movement disorders of the extremities and stereotypies can be associated with changes in the basal ganglia."<br/>"Chemotherapy: The sources describe that chemotherapy-induced polyneuropathy (CIPN) is a common side effect of various chemotherapeutic agents. This form of neuropathy can result from the toxic effects of chemotherapeutic agents on nerve fibers. Although the sources do not explicitly mention Waller degeneration as a consequence of CIPN, it is plausible that the nerve damage caused by chemotherapy can trigger degenerative processes similar to those described in Waller degeneration."</p> | <p>Movement disorders are possible clinical manifestations of basal ganglia pathologies, not the cause of basal ganglia pathologies.</p>   |
| <p>Waller degeneration after chemotherapy in CLL</p>   | <p>"Chemotherapy: The sources describe that chemotherapy-induced polyneuropathy (CIPN) is a common side effect of various chemotherapeutic agents. This form of neuropathy can result from the toxic effects of chemotherapeutic agents on nerve fibers. Although the sources do not explicitly mention Waller degeneration as a consequence of CIPN, it is plausible that the nerve damage caused by chemotherapy can trigger degenerative processes similar to those described in Waller degeneration."</p>   | <p>The LLM response incorrectly describes that chemotherapy can trigger processes similar to those in Waller degeneration, while admitting not to have found a source for this information.<br/>The lesions were described as FLAIR-hyperintense. The LLM response does not take into account that enlarged Virchow-Robin spaces would appear FLAIR-hypointense (isointense to CSF).</p> |
| <p>[Standard Prompt Introduction]<br/>Male patient in his 30s, no symptoms. The MRI shows multiple dot-like FLAIR hyperintensities grouped in the cortex of the left insula."</p>  | <p>"Enlarged Virchow-Robin spaces can sometimes appear as dot-like hyperintensities on MRI. These are perivascular spaces that can be prominent in certain individuals and are generally considered a normal variant."</p>  | <p>The LLM response does not take into account that enlarged Virchow-Robin spaces would appear FLAIR-hypointense (isointense to CSF).</p>  |



|  |  |  |
|--|--|--|
| <p>[Standard Prompt Introduction]<br/>male in his 40s, incidental clival finding in external imaging patient asymptomatic<br/>MRI features: T2 hyperintense, FLAIR iso to hypointense<br/>you are a radiologist: MRI corpus callosum, FLAIR hyperintense, diffusion restricted lenticular lesion<br/>These are some pictures of the lesion: [images]</p>   | <p>"The provided MRI images show various views of the brain, including axial, coronal, and sagittal planes."<br/>"Based on the provided MRI image and the clinical history of a female patient with vomiting and headache, the MRI findings suggest the presence of a large encapsulated peripherally enhancing lesion within the brain. This lesion is characterized by an abnormal signal intensity on the right side of the brain, which could be located in the temporal or parietal lobe. The peripheral enhancement is indicative of a rim that is likely isointense to hyperintense on T1-weighted images and hypointense on T2-weighted images, as is typical for an abscess."</p>   | <p>When asked for sample images of clival differential diagnoses, the LLM presented images of a chondrosarcoma of the humerus.<br/><br/>No sagittal images were provided.</p>  |
| <p>"you are a radiologist, female patient with vomiting, suspected brain abscess, headache. These are some MRI findings" [image]</p>   | <p>"[...] The MRI findings described include left-sided FLAIR hyperintensity, T2 hyperintensity suggestive of hemorrhage (haem hyperintense), and slight diffusion restriction in the region of the limen insulae. These imaging characteristics help narrow down the potential causes."</p>   | <p>No post-contrast images were provided. The pathology is located on the left side, not right side.</p>   |
| <p>"you are a radiologist; psychomotor retardation, dysarthria, organic psychosyndrome for 3 months, male patient in his 50s, mri findings are left side region of limen insulae, FLAIR hyperintense, T2 heme hyperintense, slight diffusion restriction, differential diagnosis"</p>  | <p>"The severe left sensorineural hearing loss and the presence of an intraventricular FLAIR isointense tumor in the cerebellum are highly suggestive of a vestibular schwannoma (acoustic neuroma), which is a benign tumor arising from the vestibular nerve sheath."</p>  | <p>"T2 heme hyperintense" does not indicate that there is hemorrhage. Hemorrhage typically shows low signal intensity in the T2 heme sequence. The LLM response suggests vestibular schwannoma as the most likely diagnosis for an intraventricular tumor in the cerebellum. However, vestibular schwannomas appear in the pontocerebellar angle, not within the ventricles.</p> |
| <p>[Standard Prompt Introduction]<br/>[severe left sensoineural heraing loss, male in his 60s]<br/>[cerebellum, intraVENTRICULAR FLAIR isointense tumor]<br/>[Standard Prompt Introduction]<br/>male in his 30s, dysarthria, dysphagia, psychomotor slowing, intracerebral lesions, suspected HIV; FLAIR hyperintense, edema, multiple lesions, ADC hyperintense, peripherally enhancing, T2 hyperintense, white matter"</p> | <p>"Arachnoid cysts can occur in the posterior fossa near the fourth ventricle. They often appear as well-defined cystic lesions with rim enhancement on MRI, similar to the described findings. However, they usually do not demonstrate restricted diffusion."<br/>"The FLAIR hyperintense lesion with contrast enhancement near the posterior horn of the left lateral ventricle, along with the sensory-motor hemiparesis on the left side, is highly suggestive of a lacunar stroke in the region of the posterior limb of the internal capsule. Lacunar strokes are small infarcts that typically occur in the deep brain structures, such as the thalamus, basal ganglia, or brainstem, and can cause pure motor, pure sensory, or mixed sensory-motor deficits."</p> | <p>Rim enhancement is highly unusual for arachnoid cysts.</p>  |
| <p>[Standard Prompt Introduction]<br/>male in his 30s old dizziness headache<br/>cystic lesion in or close to the fourth ventricle contrast enhancement at the rims diffusion restriction"</p>   | <p>Based on a lesion near the posterior horn of the left lateral ventricle, the presence of a lacunar stroke in the posterior limb of the internal capsule is suggested. These are two totally different anatomical structures.</p>  |  |

**Supplement 2: Hallucinatory LLM responses.** "Standard prompt introduction" indicates the following part of the prompt template provided to the readers: "You are a senior neuroradiologist. Below, you will find information regarding a brain MRI scan. Based on this information, identify the three most likely differential diagnoses, ranked by their likelihood. Present your findings in a table format with the following columns: 'Rank', 'Differential Diagnosis', and 'Explanation'."