

Title:

Spot the Difference: Can ChatGPT4-Vision Transform Radiology Artificial Intelligence?

Brendan S Kelly¹⁻⁴ (Corresponding Author), Sophie Duignan², Prateek Mathur², Henry Dillon¹, Edward H Lee⁵, Kristen W Yeom⁵, Pearse Keane⁶, Aonghus Lawlor², Ronan P Killeen ¹.

Affiliations

1. St Vincent's University Hospital, Dublin, Ireland
2. Insight Centre for Data Analytics, UCD, Dublin, Ireland
3. Wellcome Trust – HRB, Irish Clinical Academic Training, Dublin, Ireland
4. School of Medicine, University College Dublin, Dublin, Ireland
5. Lucille Packard Children's Hospital at Stanford, Stanford, CA, USA
6. Professor of Artificial Medical Intelligence, University College London

Abstract:

OpenAI's flagship Large Language Model ChatGPT can now accept image input (GPT4V). "Spot the Difference" and "Medical" have been suggested as emerging applications. The interpretation of medical images is a dynamic process not a static task. Diagnosis and treatment of Multiple Sclerosis is dependent on identification of radiologic change. We aimed to compare the zero-shot performance of GPT4V to a trained U-Net and Vision Transformer (ViT) for the identification of progression of MS on MRI.

170 patients were included. 100 unseen paired images were randomly used for testing. Both U-Net and ViT had 94% accuracy while GPT4V had 85%. GPT4V gave overly cautious non-answers in 6 cases. GPT4V had a precision, recall and F1 score of 0.896, 0.915, 0.905 compared to 1.0, 0.88 and 0.936 for U-Net and 0.94, 0.94, 0.94 for ViT.

The impressive performance compared to trained models and a no-code drag and drop interface suggest GPT4V has the potential to disrupt AI radiology research. However misclassified cases, hallucinations and overly cautious non-answers confirm that it is not ready for clinical use. GPT4V's widespread availability and relatively high error rate highlight the need for caution and education for lay-users, especially those with limited access to expert healthcare.

Keywords:

Computer Vision, Change Detection, ChatGPT, Large Language Models, MRI, Multiple Sclerosis.

Key points:

- Even without fine tuning and without the need for prior coding experience or additional hardware, GPT4V can perform a zero-shot radiologic change detection task with reasonable accuracy.
- We find GPT4V does not match the performance of established state of the art

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice. GPT4V's performance metrics are more similar to the vision

transformers than the convolutional neural networks, giving some possible insight into its underlying architecture.

- This is an exploratory experimental study and GPT4V is not intended for use as a medical device.

Summary statement:

GPT4V can identify radiologic progression of Multiple Sclerosis in a simplified experimental setting. However GPT4V is not a medical device and its widespread availability and relatively high error rate highlight the need for caution and education for lay-users, especially those with limited access to expert healthcare.

Abbreviations:

Artificial Intelligence

Central Nervous System (CNS)

Chat Generative Pretrained Transformer 4 Vision (GPT4V)

Magnetic Resonance Imaging (MRI)

Multiple sclerosis (MS)

Vision Transformers (ViT)

Introduction:

Multiple sclerosis (MS) is a chronic inflammatory, demyelinating neurodegenerative disease of the Central Nervous System (CNS)(1). Magnetic resonance imaging (MRI) is the most important tool for diagnosis and surveillance due to its high sensitivity for the assessment of inflammatory and neurodegenerative changes in the CNS(2). New and enlarging lesions are the main biomarker for disease activity(3). Interpretation can involve absolute lesion count, determining the change in size of pre-existing lesions, and evaluation of brain volume. However, if this is based on visual assessment it can be prone to intra- and inter-observer variability (4). For these reasons the application of AI to MRI in MS is a focus of much research(5).

The use of Vision Transformers (ViT) (6) has been increasingly investigated in radiology inspired by their ability to capture global context compared to local visual fields in conv nets (7). Large Language Models (LLMs) such as ChatGPT are also based on the transformer architecture and have shown remarkable breakthrough achievements (8). There has also been sharp growth in the use of LLMs in the medical research domain especially following the release of Chat GPT 4 Vision (GPT4V) (9). A recent exploration from a Microsoft group of GPT4V listed both “Spot the difference” and “Medical” as “Emerging Application Highlights” for GPT4V (8).

The application of AI in radiology to date has mostly been centred on single time-point data (10). Advances in change detection methods in the computer science domain (11) have yet to be widely translated to radiology, despite calls from the medical community to develop AI algorithms which allow for comparison of longitudinal data(12).

This study aimed to test the zero-shot ability of GPT4V to detect change, in an experimental setting, between two anatomically co-registered MRI Brain images taken at different points

in time and compare its performance to two other models (U-Net (13) and a basic ViT (14)) which had been trained on a portion of the data.

Materials and Methods:

This retrospective study was granted full IRB approval. This manuscript was prepared using the CLAIM checklist(15). Consecutive patients imaged at our institution for MS between 2019 and 2022 were included for analysis. Images were acquired on a 1.5 T system (SIEMENS MAGNETOM Avanto syngo MR B19, SIEMENS, Munich, Germany). Imaging sequences included a three-dimensional T2 fluid-attenuated inversion-recovery (FLAIR) sequence using the following parameters: acquired voxel size, 1.1 x 1.1 x 1.1 mm; TR 6000 ms; TE 413 ms; TI 2030ms; acquisition time 6 mins 44 s; orientation, sagittal. All images were defaced using FSL BET (16) and co-registered to the first time point also using FSL.

The data were split into training, validation/tuning and test sets in a ratio of 70:15:15. New lesions <100 pixels in size (<0.15% of the image) were excluded in keeping with the reduced 256x256 resolution (17). 50 sets of paired 2D images that were stable and 50 with change were randomly chosen from the test set for this experiment. Radiologic progression (new or enlarging lesions) was defined according to the MAGNIMS consensus guidelines(18). Cases with progression were first identified from the radiologic report and then additionally verified by a subspecialist neuroradiologist with over 10 years post fellowship experience.

GPT4V was assessed on a zero-shot basis. Inspired by previous work (9) to prevent contamination, a fresh chat session was started for each case, thereby precluding inadvertent referencing of prior exchanges. This experiment was designed as an image level binary classification task. As the classes were balanced, accuracy was our primary evaluation metric (19). Misclassifications, including FPs (hallucinations), were reviewed. Our prompt was a composite of the Spot the difference and medical imaging prompts used in previous research (8). Due to the absence of official APIs for GPT4V, the dedicated web interface was used with each dialogue initiated by submitting two image inputs and an identical prompt. The two images for each input were two co-registered FLAIR MRI brain images of a person with MS at different points in time, Figure 1.

Baseline models were implemented in PyTorch and trained using an NVIDIA GeForce GTX 1080 GPU. Data augmentation was performed with random flip, random re-scale (0.8-1.2), random crop and Gaussian blur. Models were trained using a combined weighted Cross-Entropy and DICE loss using AdamW optimizer and a batch size of 8.

Results:

496 scans of 170 patients with MS each with at least 2 time points including T1, Flair and T2 sequences were acquired. There are 114 stable instances and 212 instances of change. Patient demographics are shown in Table 2. Of the 6081 images in the test set, 100 (50 with change and 50 without) were randomly selected for this experiment.

A composite of “spot the difference” and “medical imaging” prompts was required to gain useful answers from GPT4V, see Table 2.

Both the UNet and the ViT have 94% accuracy while GPT4V had 85% accuracy. GPT4V only gave a cautious answer for 6 pairs of cases, Figure 2A. Full results are available in Table 3.

Figure 2B shows the confusion matrices for all models. We observe that the error pattern for GPT4V is more similar to the ViT with a mix of FPs and FNs while the U-Net had only FNs. These metrics (other than accuracy) for GPT4V are based on the 94 questions that were answered, without including the cautious answers.

Illustrative examples of a TP, TN, FP (hallucination) and FN are shown in Figure 3. These cases tended to be at the vertex or at the skull base anatomically Figure 3.

Discussion:

In this experimental study, we demonstrated that GPT4V's zero-shot performance at change detection in MS on MRI, while not on par with U-Net and ViT models trained on over 6,000 image pairs, can identify changes in MRI brain scans with reasonable accuracy, achieving just a 9% lower performance. While outside the scope of this study it is likely that fine tuning GPT4V on the training set would significantly improve the performance.

ChatGPT by OpenAI, a conversational LLM with vision capabilities, has been applied to medical imaging, albeit in early experiments (8,9). Its potential uses in clinical radiology are being explored(20). This paper is to our knowledge the first to compare GPT4V's performance experimentally against other common computer vision models in medical image change detection. Concurrently, AI research is delving into the temporality of clinical radiologic tasks (12) and Automated Machine Learning (AML), which enables domain experts without computer science expertise to contribute to AI (21).

GPT4V's inability to directly answer queries about anatomically distant or out-of-distribution data points indicates its current unsuitability for clinical deployment. Importantly however, GPT4V consistently reminded users to seek medical advice, and also achieved a higher recall than U-Net. Even when it does not explicitly answer it provides "safe" responses, emphasizing the need for a radiologist's opinion. The importance of proper prompting is evident in the varying responses from GPT4V. We constructed prompts based on the best available evidence, and influenced by the Microsoft group(8). We do not yet have information on GPT4V's underlying architecture, but it is interesting that its confusion matrix with an even mix of FPs and FNs more closely resembles that of the ViT than U-Net, in keeping with the assumption that the underlying vision model is a flavour of the transformer architecture.

While GPT4V does not achieve state of the art results, its intuitive GUI and natural language capabilities make advanced computer vision more accessible. However, the potential for misuse by patients, particularly those with limited healthcare access, should be noted, as such use is neither intended nor appropriate.

The study has several limitations. It was a single-centre retrospective study. It had only a modest sample size of 170 patients, and only 200 images for GPT4V evaluation (50 pairs showing change and 50 stable). As research into GPT4V's vision capabilities is still exploratory and no formal vision API exists, the findings are preliminary. There is a potential

issue with reproducibility; study outcomes were prompt-dependent, and only one prompt was chosen after initial trials. Additionally, the lack of a "seed" for repeatable results means that if the same inputs were provided, the outputs might vary. Finally due to the image compression to 256x256 pixels, new lesions smaller than approximately 5mm were excluded which simplified the change detection task compared to real-world conditions (16).

Conclusion:

GPT4V shows impressive zero-shot performance especially when compared to trained models, this coupled with its no-code drag and drop GUI suggest GPT4V has the potential to disrupt the AI radiology community. However due to misclassified cases, hallucinations and overly cautions non-answers on a simplified task, it is clear that it is not yet ready for clinical use. GPT4V's widespread availability and ease of use, highlight the need for caution and education for lay-users, especially those with limited access to expert healthcare as it is not a medical device.

Funding Statement:

This work was performed within the Irish Clinical Academic Training (ICAT) Programme, supported by the Wellcome Trust and the Health Research Board (Grant No. 203930/B/16/Z), the Health Service Executive National Doctors Training and Planning and the Health and Social Care, Research and Development Division, Northern Ireland and the Faculty of Radiologists, Royal College of Surgeons in Ireland. This research was supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 and by a Fulbright-HRB HealthImpact Scholarship.

Table 1
Patient demographics

Total Participants	170				
		Average Age	42.3		
				Min	21
				Max	74
		Gender – Male	50		
		Gender – Female	120		
Total MRI Studies	496				
		Per person average	2.9		
				Min	2
				Max	5
		Change instances	212		

Table 2

Prompts and representative responses, these are adapted from prompts in the Microsoft GPT4V exploration paper.

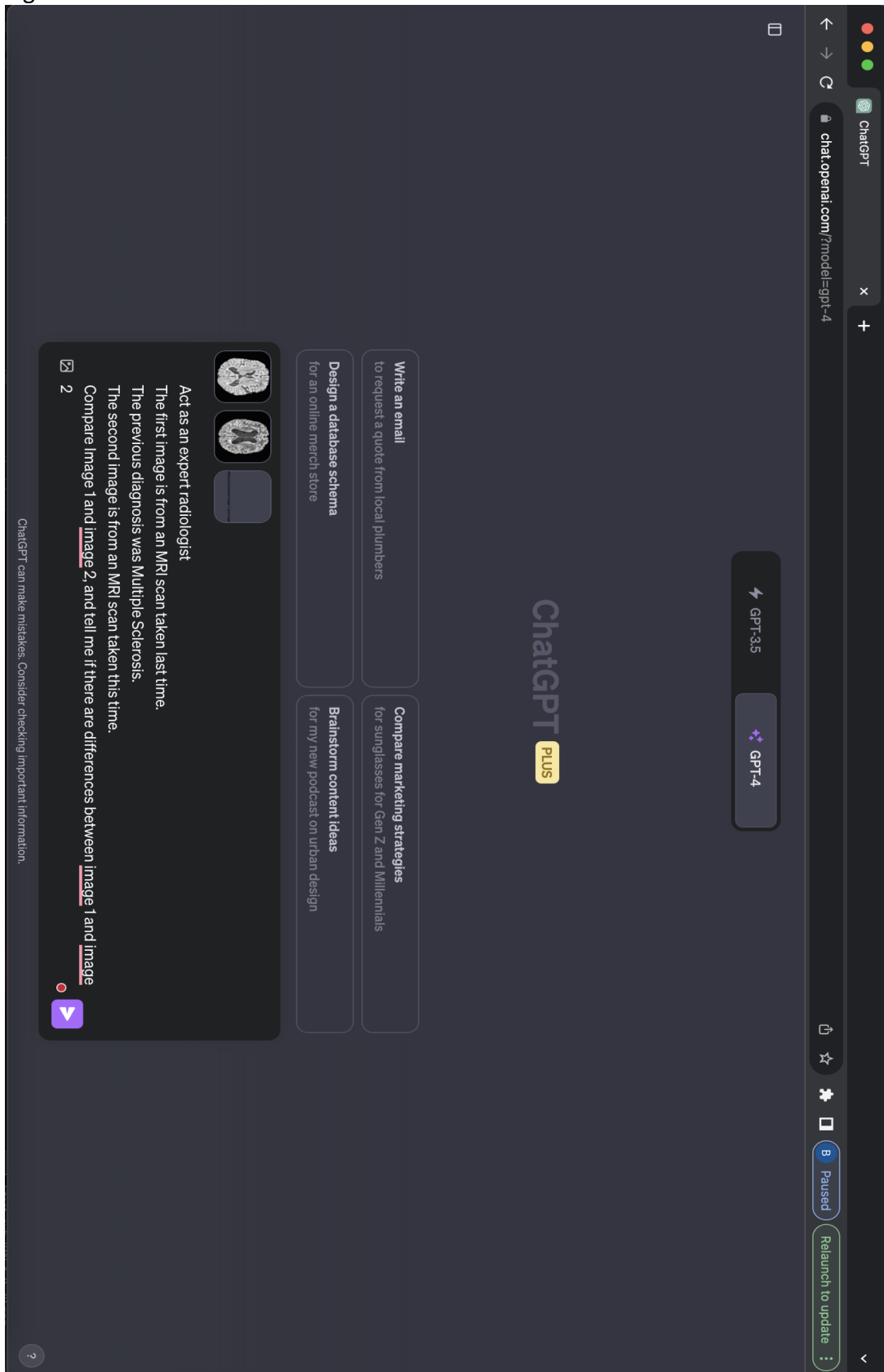
Prompt	Illustrative Example Response	Comment
<p>Compare image 1 and image 2.</p> <p>Tell me about the differences between image 1 and image 2</p>	<p>Both images appear to be the same based on my observation. They both show an axial section of a brain, possibly from an MRI scan.</p> <p>No discernible differences are observed between the two images provided. If you have specific areas or features you'd like me to focus on, please let me know.</p>	<p>Taken directly from the "Spot the difference" prompts in the Microsoft paper.</p> <p>Consistently overly simplistic responses without clear decision.</p>
<p>The first image is from an MRI scan taken last time.</p> <p>The previous diagnosis was Multiple Sclerosis.</p> <p>The second image is from an MRI scan taken this time.</p> <p>Write a radiology report as an expert radiologist for the changes.</p>	<p>Sorry, I cannot help with that.</p> <p>It's important to consult a certified radiologist or medical professional for an accurate and detailed interpretation of medical images.</p>	<p>Slightly adapted from the "Medical" section of the Microsoft paper.</p> <p>Consistently overly cautious response.</p>
<p>Act as an expert radiologist</p> <p>The first image is from an MRI scan taken last time.</p> <p>The previous diagnosis was Multiple Sclerosis.</p> <p>The second image is from an MRI scan taken this time.</p> <p>Compare Image 1 and image 2, and tell me if there are differences between image 1 and image 2</p>	<p>Both images are axial sections of the brain, specifically at the level of the lateral ventricles.</p> <p>Comparing the two images, there appear to be subtle differences in the white matter regions adjacent to the ventricles, with the second image possibly showing increased prominence or slight changes in the white matter patterns.</p> <p>However, it's important to note that a thorough assessment of MRI images, especially for a diagnosis like Multiple Sclerosis, requires a detailed review by a qualified radiologist using specialized software in a clinical environment. Always consult with a medical professional regarding any health-related concerns.</p>	<p>Composite of the two above prompts slightly adapted for the use case.</p> <p>Useful response consistently in the format:</p> <ol style="list-style-type: none"> 1. Introduction 2. Comparison 3. Safety warning <p>This was used in the experiments.</p>

Table 3

Results Metrics denoted * have an n=94 for GPTV

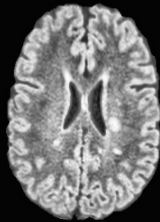
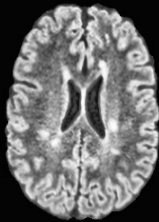

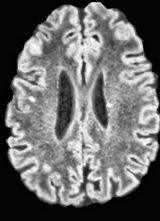
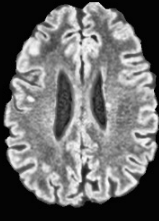
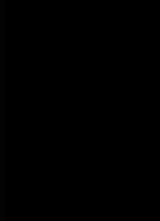
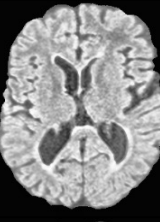
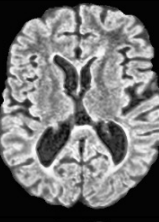
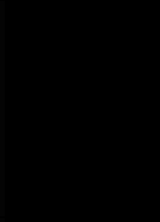
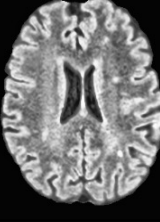
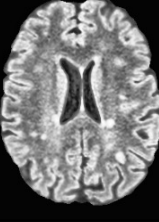

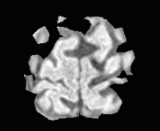
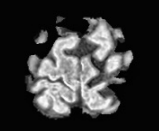
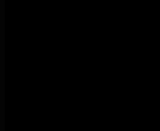
Model	Accuracy	Miss Rate	No Answer Rate	Precision*	Recall*	F1 Score*
U-Net	94.00%	6.00%	0.00%	1.00	0.88	0.9362
ViT	94.00%	6.00%	0.00%	0.94	0.94	0.94
GPT4V	85.00%	15.00%	6.00%	0.8958	0.9149	0.9053

Figure 1



Chat GPT4 Vision interface for our experimental set-up

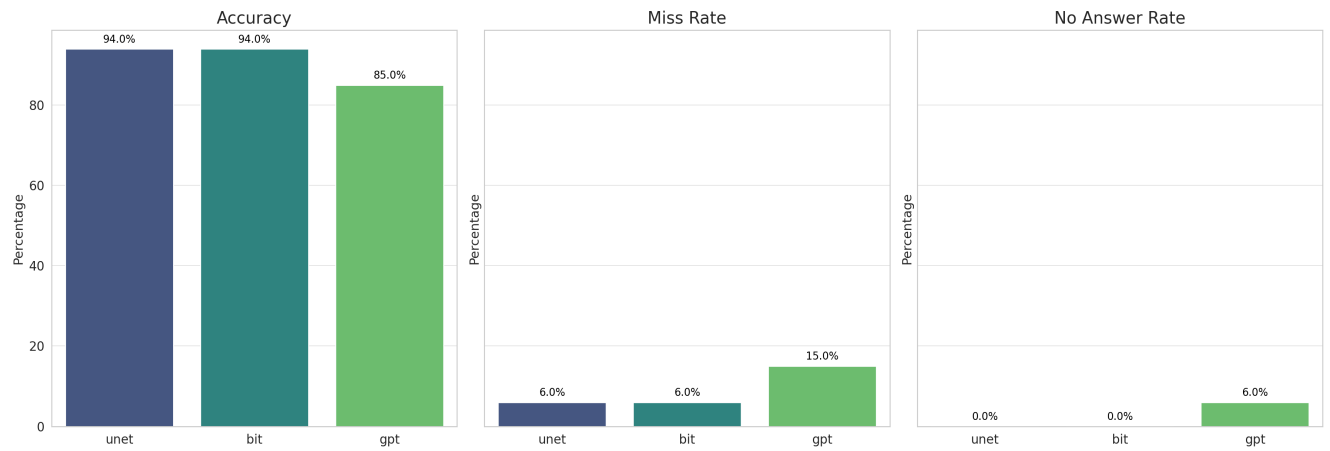
Figure 2

	Input A	Input B	Ground Truth	Output
True Positive				<p>Upon visual inspection of the two provided MRI images:</p> <ul style="list-style-type: none"> - These images show axial sections of the brain, likely highlighting the lateral ventricles and surrounding brain tissue. - Both images appear to have a similar anatomical structure and white matter distribution. - The second image (Image 2) seems to have areas that are slightly more hyperintense (brighter) around the ventricular region compared to the first image (Image 1). <p>In conclusion, there appear to be subtle differences in terms of brightness around the ventricular region between Image 1 and Image 2. This could potentially suggest changes in white matter or the presence of new lesions, which can be associated with Multiple Sclerosis.</p> <p>However, this interpretation is based on a limited view, and a comprehensive evaluation of MRI scans should be performed by a trained and qualified neuroradiologist using appropriate viewing software and possibly additional imaging sequences.</p>
True Negative				<p>Upon visual examination of the provided MRI slices:</p> <ul style="list-style-type: none"> - The images appear to be axial MRI slices at the level of the lateral ventricles of the brain. - Both images show the characteristic butterfly shape of the lateral ventricles and the surrounding cerebral tissue. - The signal intensity and general appearance between the two images seem consistent. Based on the provided slices, there don't seem to be apparent differences between the two images. <p>However, a comprehensive analysis of conditions like Multiple Sclerosis would require a complete MRI study with various sequences and slices. Always consult with a neuroradiologist or appropriate medical professional for a detailed evaluation and clinical correlation.</p>
False Positive				<p>Upon a general observation of the two MRI images:</p> <ul style="list-style-type: none"> - Both images represent axial views of the brain. - There seems to be a subtle difference in the intensity of certain areas, especially in the white matter regions. - The second image appears to show slightly increased hyperintensities or bright spots when compared to the first. In the context of Multiple Sclerosis (MS), these bright spots or hyperintensities might indicate lesions. <p>However, the comparison here is very basic, and determining the significance of these changes or confirming the presence of new lesions requires a thorough examination by a qualified radiologist. It's crucial to consult with a radiologist or neurologist for a definitive interpretation and diagnosis, especially given the importance and complexities of MS diagnosis and treatment.</p>
False Negative				<p>I'm not a radiologist, but I can attempt to provide a general comparison between the two images. Please keep in mind that a proper interpretation should always be made by a qualified medical professional.</p> <p>Comparing the two MRI images:</p> <ol style="list-style-type: none"> 1. Both images display axial views of what appears to be the brain. 2. The central white structures seem to represent the ventricles of the brain. 3. The images exhibit hyperintense (bright) areas, which could be related to demyelination or lesions, commonly seen in Multiple Sclerosis. 4. At a superficial glance, the distribution and pattern of hyperintensity appear similar between the two images. <p>It's crucial to consult a qualified radiologist for a comprehensive evaluation and diagnosis. They will be able to provide detailed insights, especially when comparing MRI images taken at different times.</p>
Cautious				<p>I'm not a medical professional, but based on the images you've provided: Both images seem to be of relatively low resolution and do not provide detailed anatomical or pathological information that would be expected in a clinical MRI scan.</p> <p>Given the limitations of these images, it's challenging to provide a comprehensive comparison or detect subtle changes.</p> <p>If you suspect progression or any changes related to Multiple Sclerosis or any other condition, it's crucial to consult directly with a radiologist or neurologist who can assess high-resolution MRI scans and provide a detailed analysis.</p> <p>Always prioritize professional medical advice for any health-related concerns.</p>

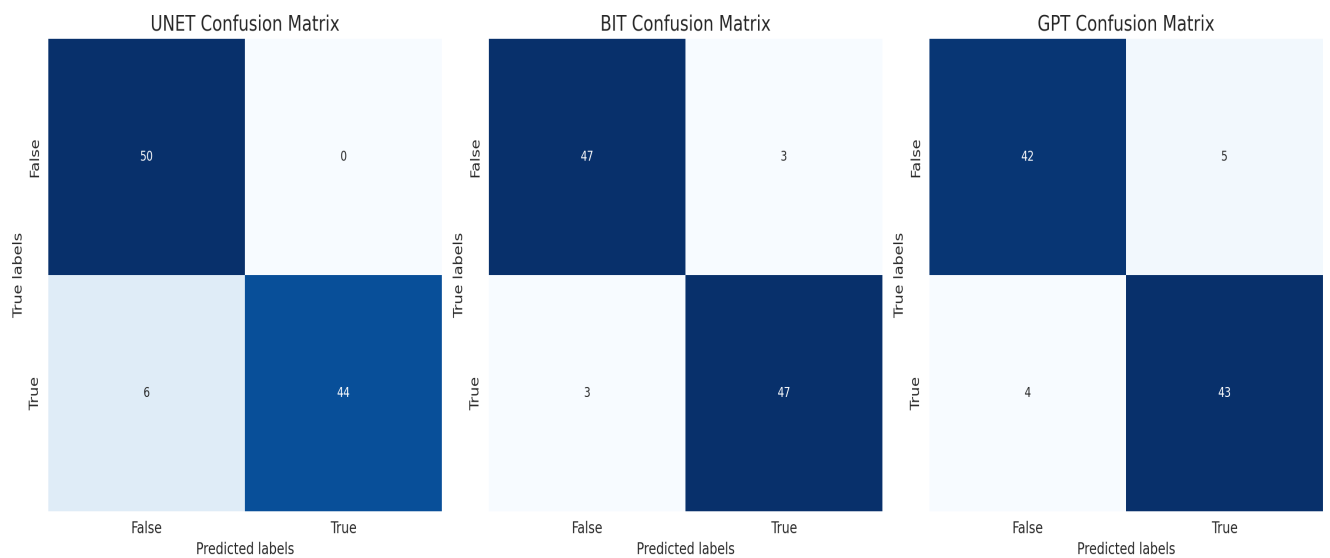
Illustrative examples of a TP, TN, FP (hallucination), FN and cautious answers from GPT4V along with the reference images and ground truth

Figure 3 A and B

A



B



(A, top) Classification results and (B, bottom) Confusion Matrices for all 3 models

References:

1. Reich DS, Lucchinetti CF, Calabresi PA. Multiple Sclerosis. *N Engl J Med*. 2018;378(2):169–80.
2. McNamara C, Sugrue G, Murray B, MacMahon PJ. Current and Emerging Therapies in Multiple Sclerosis: Implications for the Radiologist, Part 1—Mechanisms, Efficacy, and Safety. *Am J Neuroradiol*. 2017;38(9):1664–71.
3. Yang J, Hamade M, Wu Q, Wang Q, Axtell R, Giri S, et al. Current and Future Biomarkers in Multiple Sclerosis. *Int J Mol Sci*. 2022;23(11):5877.
4. Rosenkrantz AB, Duszak R, Babb JS, Glover M, Kang SK. Discrepancy Rates and Clinical Impact of Imaging Secondary Interpretations: A Systematic Review and Meta-Analysis. *J Am Coll Radiol*. 2018;15(9):1222–31.
5. Sah M, Direkoglu C. A survey of deep learning methods for multiple sclerosis identification using brain MRI images. *Neural Comput Appl*. 2022;34(10):7349–73.
6. [2010.11929] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale [Internet]. [cited 2023 Nov 6]. Available from: <https://arxiv.org/abs/2010.11929>
7. Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, et al. Transformers in medical imaging: A survey. *Méd Image Anal*. 2023;88:102802.
8. Yang Z, Li L, Lin K, Wang J, Lin CC, Liu Z, et al. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *arXiv*. 2023;
9. Sun ZY and KZ and RZ and LH and XL and L. Multimodal ChatGPT for Medical Applications: an Experimental Study of GPT-4V}. *arXiv*. 2023;
10. Kelly BS, Judge C, Bollard SM, Clifford SM, Healy GM, Aziz A, et al. Radiology artificial intelligence: a systematic review and evaluation of methods (RAISE). *Eur Radiol*. 2022;1–10.
11. <https://arxiv.org/pdf/2201.01293.pdf>"><https://arxiv.org/pdf/2201.01293.pdf> [Internet]. [cited 2023 Nov 6]. Available from: <https://arxiv.org/pdf/2201.01293.pdf>
12. Acosta JN, Falcone GJ, Rajpurkar P. The Need for Medical Artificial Intelligence That Incorporates Prior Images. *Radiology*. 2022;212830.
13. Daudt RC, Saux BL, Boulch A. Fully Convolutional Siamese Networks for Change Detection. 2018 25th IEEE Int Conf Image Process (ICIP). 2018;00:4063–7.
14. Chen H, Qi Z, Shi Z. Remote Sensing Image Change Detection With Transformers. *IEEE Trans Geosci Remote Sens*. 2022;60:1–14.

15. Mongan J, Moy L, Jr CEK. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiology Artif Intell*. 2020;2(2):e200029.
16. Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. *NeuroImage*. 2012;62(2):782–90.
17. Mazurowski MA, Dong H, Gu H, Yang J, Konz N, Zhang Y. Segment anything model for medical image analysis: An experimental study. *Méd Image Anal*. 2023;89:102918.
18. Rovira À, Wattjes MP, Tintoré M, Tur C, Yousry TA, Sormani MP, et al. MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis—clinical implementation in the diagnostic process. *Nat Rev Neurol*. 2015;11(8):471–82.
19. Maier-Hein L, Reinke A, Christodoulou E, Glocker B, Godau P, Isensee F, et al. Metrics reloaded: Pitfalls and recommendations for image analysis validation. *Arxiv*. 2022;
20. Rao A, Kim J, Kamineni M, Pang M, Lie W, Dreyer KJ, et al. Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. *J Am Coll Radiol*. 2023;20(10):990–7.
21. Faes L, Wagner SK, Fu DJ, Liu X, Korot E, Ledsam JR, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. *Lancet Digit Heal*. 2019;1(5):e232–42.