Original Paper

Using Natural Language Processing (GPT-4) for Computed Tomography Image Analysis of Cerebral Hemorrhages in Radiology: Retrospective Analysis

Daiwen Zhang[1,2*], PhD; Zixuan Ma[1,2*], PhD; Ru Gong[1*], PhD; Liangliang Lian[3*], MD; Yanzhuo Li[4*], MD; Zhenghui He[1,2], PhD; Yuhan Han[1,2], PhD; Jiyuan Hui[1], PhD; Jialin Huang[2], PhD; Jiyao Jiang[1,2], PhD; Weiji Weng[1,2*], PhD; Junfeng Feng[1,2], PhD

[1]Brain Injury Centre, Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[2]Shanghai Institute of Head Trauma, Shanghai, China
[3]Department of Radiology, Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[4]Department of Radiology, Minhang Hospital, Fudan University, Shanghai, China
[*]these authors contributed equally

Corresponding Author:
Junfeng Feng, PhD
Brain Injury Centre
Ren Ji Hospital
Shanghai Jiao Tong University School of Medicine
160 Pujian Road, Pudong New District
Shanghai, 200127
China
Phone: 86 136 1186 0825
Fax: 86 021 68383709
Email: fengjfmail@163.com

Abstract

Background: Cerebral hemorrhage is a critical medical condition that necessitates a rapid and precise diagnosis for timely medical intervention, including emergency operation. Computed tomography (CT) is essential for identifying cerebral hemorrhage, but its effectiveness is limited by the availability of experienced radiologists, especially in resource-constrained regions or when shorthanded during holidays or at night. Despite advancements in artificial intelligence–driven diagnostic tools, most require technical expertise. This poses a challenge for widespread adoption in radiological imaging. The introduction of advanced natural language processing (NLP) models such as GPT-4, which can annotate and analyze images without extensive algorithmic training, offers a potential solution.

Objective: This study investigates GPT-4’s capability to identify and annotate cerebral hemorrhages in cranial CT scans. It represents a novel application of NLP models in radiological imaging.

Methods: In this retrospective analysis, we collected 208 CT scans with 6 types of cerebral hemorrhages at Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine, between January and September 2023. All CT images were mixed together and sequentially numbered, so each CT image had its own corresponding number. A random sequence from 1 to 208 was generated, and all CT images were inputted into GPT-4 for analysis in the order of the random sequence. The outputs were subsequently examined using Photoshop and evaluated by experienced radiologists on a 4-point scale to assess identification completeness, accuracy, and success.

Results: The overall identification completeness percentage for the 6 types of cerebral hemorrhages was 72.6% (SD 18.6%). Specifically, GPT-4 achieved higher identification completeness in epidural and intraparenchymal hemorrhages (89.0%, SD 19.1% and 86.9%, SD 17.7%, respectively), yet its identification completeness percentage in chronic subdural hemorrhages was very low (37.3%, SD 37.5%). The misidentification percentages for complex hemorrhages (54.0%, SD 28.0%), epidural hemorrhages (50.2%, SD 22.7%), and subarachnoid hemorrhages (50.5%, SD 29.2%) were relatively high, whereas they were relatively low for acute subdural hemorrhages (32.6%, SD 26.3%), chronic subdural hemorrhages (40.3%, SD 27.2%), and intraparenchymal hemorrhages (26.2%, SD 23.8%). The identification completeness percentages in both massive and minor bleeding showed no significant difference (*P*=.06). However, the misidentification percentage in recognizing massive bleeding was significantly lower than that for minor bleeding (*P*=.04). The identification completeness percentages and misidentification percentages for cerebral hemorrhages at different locations showed no significant differences (all *P*>.05). Lastly, radiologists showed relative acceptance regarding identification completeness (3.60, SD 0.54), accuracy (3.30, SD 0.65), and success (3.38, SD 0.64).

Conclusions: GPT-4, a standout among NLP models, exhibits both promising capabilities and certain limitations in the realm of radiological imaging, particularly when it comes to identifying cerebral hemorrhages in CT scans. This opens up new directions and insights for the future development of NLP models in radiology.

Trial Registration: ClinicalTrials.gov NCT06230419; https://clinicaltrials.gov/study/NCT06230419

(J Med Internet Res 2024;26:e58741) doi: 10.2196/58741

KEYWORDS

GPT-4; natural language processing; NLP; artificial intelligence; AI; cerebral hemorrhage; computed tomography; CT

Introduction

Cerebral hemorrhage mainly encompasses intracranial bleeding resultant from trauma, hypertension, or cerebral vascular disorders. Typically, it manifests acutely, with severe symptoms and a prognosis that is often unfavorable, thereby imposing substantial economic burdens on individuals, families, and society [1-3]. Hence, the prompt and accurate diagnosis of cerebral hemorrhage is of paramount clinical and societal importance.

Computed tomography (CT) scanning stands as the quintessential diagnostic tool for cerebral hemorrhage. An expedient and precise diagnosis not only facilitates the comprehensive evaluation of patient conditions, the optimization of therapeutic strategies, and the prognostication of outcomes but also enhances communication between clinicians and patients or their respective families. Nonetheless, several challenges persist. First, CT diagnosis of cerebral hemorrhage often requires experienced radiologists. In low-income areas or countries, there is a significant shortage of these specialists [4,5]. Radiologists lacking sufficient experience will lead to missed or incorrect diagnosis, which can be potentially fatal in the clinical management of cerebral hemorrhage. Second, CT diagnosis of cerebral hemorrhage requires rapid reporting, but even in high-income areas and countries, the substantial workload makes it difficult for radiologists to provide prompt and comprehensive radiology reports [6]. Third, cultivating an experienced radiologist requires a mature medical education system and a substantial database of imaging resources, which is currently lacking in China [7].

Artificial intelligence (AI) has emerged as a formidable instrument in computer-aided diagnosis across various medical fields, enhancing classification [8], detection [9], and image segmentation [10]. AI proves particularly beneficial in supporting decision-making and rapid diagnosis, especially in underserved rural areas or densely populated regions with a scarcity of medical imaging experts [11]. Moreover, AI has substantially advanced medical education [12-14].

Currently, various AI models, including You Only Look Once and bespoke convolutional neural networks [15-17], have demonstrated efficacy in detecting cerebral hemorrhages in CT imagery. However, the construction and utilization of these AI models necessitate an understanding of computer algorithms. For most radiologists, the implementation of these models poses a significant challenge, indicating a need for more intuitive and accessible AI-driven diagnostic solutions.

The recent introduction of new natural language processing (NLP) models, such as ChatGPT (OpenAI) and the latest iteration of GPT-4, has enabled imaging analysis capabilities without the need for intricate algorithmic knowledge and extensive training [18,19]. Previously, large language models faced the challenge of hallucination, presenting presumably persuasive results that were not based on the input [20,21]. However, many studies suggest that GPT-4 can assist in generating, extracting, and interpreting radiology reports [22-24]. GPT-4 can play a role in the diagnosis and treatment of oral diseases, orthopedic diseases, neurological disorders, and pulmonary diseases using radiology reports [23,25-27]. GPT-4 is capable of accurately extracting lesion information from radiology reports, such as identifying metastatic diseases and generating correct labels for tumor progression [23,24]; it can integrate radiology reports and medical history to make diagnoses [27]; and it can even provide some treatment suggestions based on the radiology reports [26]. However, current research focuses only on GPT-4's capabilities with free text, and its potential for imaging recognition has not yet been explored.

This work aims to delineate our study by utilizing GPT-4 for the identification and annotation of cerebral hemorrhages in CT images, including various types of intracranial bleeding. It is the first attempt to directly use GPT-4 to identify CT images of cerebral hemorrhages. It broadens the horizon for GPT-4 applications in the medical field, offering fresh perspectives for the development of GPT-4 and other NLP models in health care.

Methods

Ethical Considerations

This retrospective study was registered at ClinicalTrials.gov (NCT06230419) and approved by the Ethics Committee of Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine (IIT-2024-0006). This study exclusively utilized CT images, which do not include personally identifiable information or sensitive individual data. There was no direct interaction with patients, complying with the principles of ethical conduct in research. Meanwhile, the Ethics Committee of Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine, approved the usage of CT images without any identifiable private information. All the data are securely stored and only accessed by personnel involved in the research team.

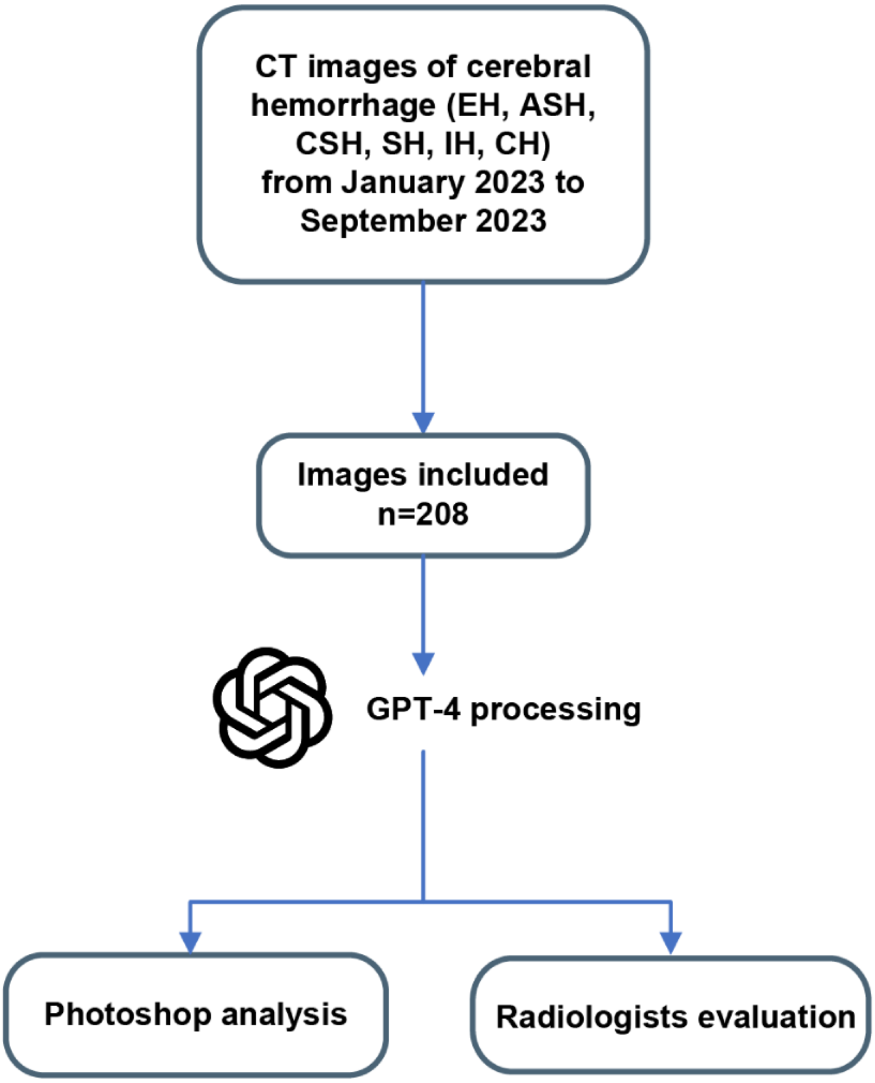
Study Design

As shown in Figure 1, raw CT images of different types of cerebral hemorrhages were collected between January and September 2023 from the radiology database of Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine. Since GPT-4 cannot recognize continuous CT images, we first preprocessed the CT images. We chose the horizontal cranial CT image with the largest volume of hemorrhage in the brain window (window width: 90, window level: 35) as the representative image. Representative CT images were exported to be unified as JPG format files with a size of 700×700 pixels to minimize the influence from the image format and the image size.

By employing a predefined question-and-answer mode, we guided GPT-4 to annotate the hemorrhagic lesions in the provided images (Multimedia Appendix 1). All CT images of different types of cerebral hemorrhages were mixed together and each CT image was sequentially numbered, so all 208 CT images had their own corresponding number. A random sequence from 1 to 208 was generated using Python (version 3.8; Python Software Foundation), and all CT images were inputted into GPT-4 for analysis in the order of the random sequence. Once GPT-4 finished the analyze process, all annotated pictures were downloaded and saved as JPG format files.

All processed CT images were both analyzed by Photoshop (version 24.3.0; Adobe) and given to 10 radiologists to evaluate identification completeness, accuracy, and success.

Figure 1. Flowchart of the study design. ASH: acute subdural hemorrhage; CH: complex hemorrhage; CSH: chronic subdural hemorrhage; CT: computed tomography; EH: epidural hemorrhage; IH: intraparenchymal hemorrhage; SH: subarachnoid hemorrhage.


CT images of cerebral hemorrhage (EH, ASH, CSH, SH, IH, CH) from January 2023 to September 2023
Images included n=208
GPT-4 processing
Photoshop analysis
Radiologists evaluation

GPT-4 Update

We utilized the latest updated GPT-4 released by OpenAI in November 2023, which at present, empowers more powerful image analysis in GPT-4.

Photoshop Analysis

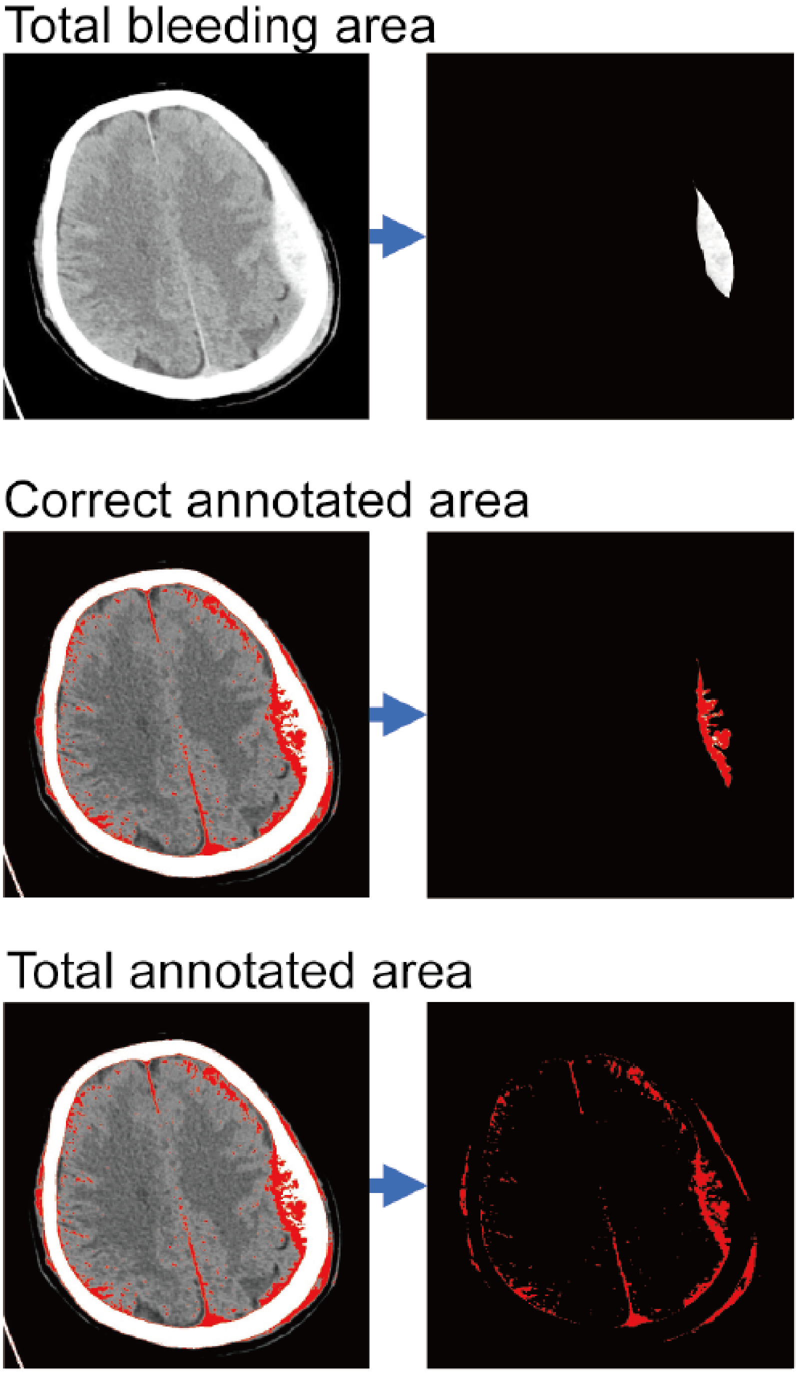
After all the processed images were collected, the hemorrhage areas and the annotation areas were isolated from the images using Photoshop [28]. According to previous studies, the ABC/2 method has been universally applied in clinical medicine to estimate cerebral hemorrhage volumes swiftly [29,30]. The same method was used in the calculation of intraparenchymal hemorrhage volumes, given that intraparenchymal hemorrhages typically conform to the assumption of this formula: the hemorrhages approximate an elliptical shape. After tallying all intraparenchymal hemorrhage volumes, based on the corresponding bleeding areas measured using Photoshop, the hemorrhage volume per pixel was determined and the hemorrhage volume for each CT image in the other 5 types of irregular cerebral hemorrhages was calculated.

After the size of the exact image's bleeding and annotated part were counted, the identification completeness and misidentification percentages were calculated through the following formulas (Figure 2):

Identification completeness percentage = correct annotated area / total bleeding area × 100%

Misidentification percentage = (total annotated area – correct annotated area) / total annotated area × 100%

Figure 2. Calculation methods of identification completeness and misidentification percentage using Photoshop. The 3 pairs of pictures show the method of using Photoshop to isolate the total bleeding area, the correct annotated area, and the total annotated area.



Radiologist Evaluation

For the purpose of evaluating the competence of GPT-4 in cerebral hemorrhage identification, 10 professional radiologists were invited to compare the original CT images and the annotated images. We designed a 4-point scale questionnaire. In this questionnaire, the original image and the annotated image were compared in pairs, and those pairs of CT images were shown to the radiologists. To avoid ambiguous and unclear outcomes, the choices of the questionnaire did not contain neutral responses. The questionnaire was completed by radiologists separately and alone. The criteria and groups used for rating are shown in Multimedia Appendix 2.

Statistical Analysis

The data in this study were all continuous and conformed to normal distribution and homogeneity of variance; therefore, we chose the 2-tailed Student *t* test and 1-way ANOVA for statistical analysis. SPSS (version 16.0; IBM Corp) was used to perform all statistical analysis. Summary data are presented as mean and SD, with statistical significance assessed by 2-tailed Student *t* test for 2-group comparisons or 1-way ANOVA for comparisons with more than 2 groups. *P*<.05 was considered statistically significant.

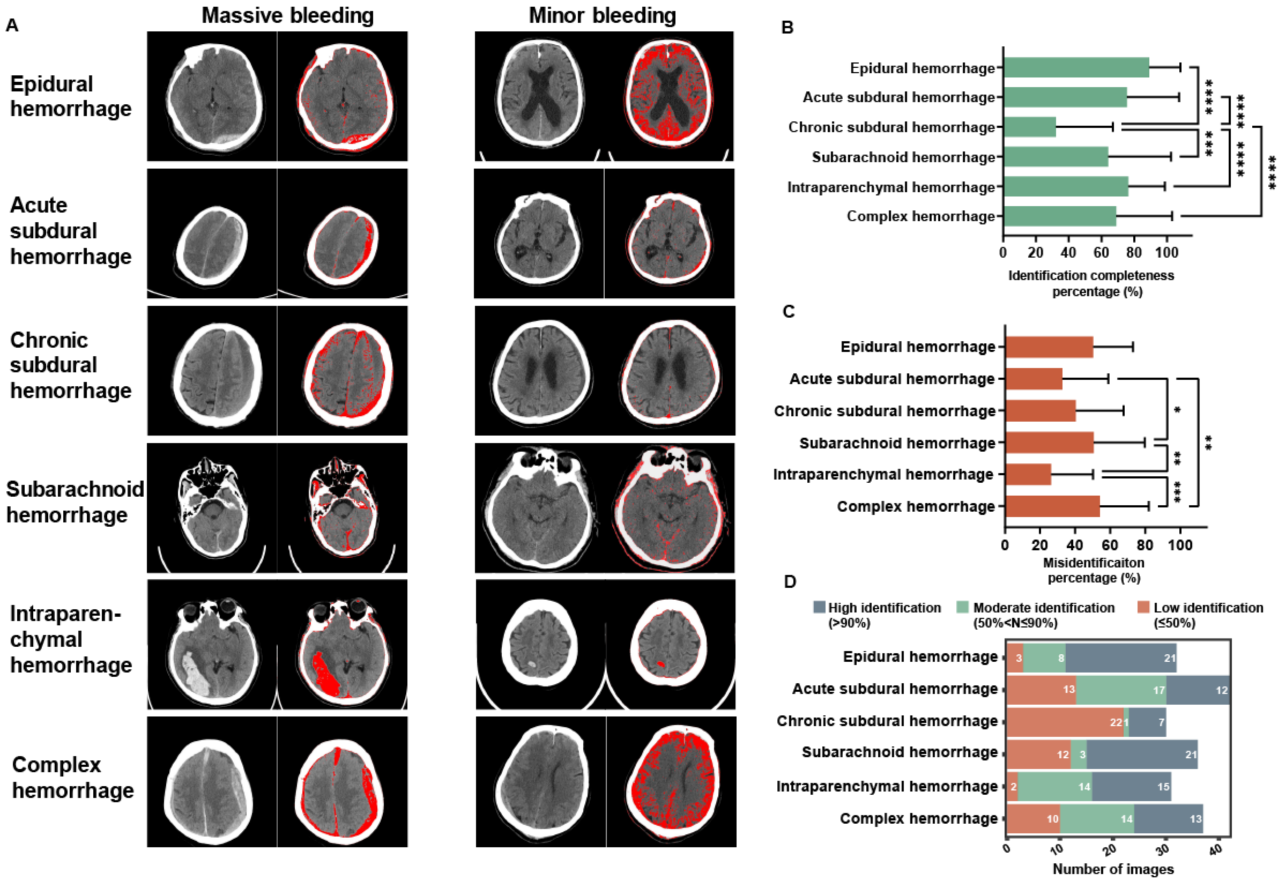
Results

A total of 208 CT scans of different types of cerebral hemorrhages were collected between January and September in 2023 from the radiology database. These images consisted of epidural hematomas (32 images); acute subdural hematomas (42 images); chronic subdural hematomas (30 images); subarachnoid hemorrhages (36 images); intraparenchymal hemorrhages (31 images); and complex hemorrhages, which included 2 or more different types of intracranial bleeding (37 images).

First and foremost, we were keen to assess whether GPT-4 could identify cerebral hemorrhages and the completeness of such detections. In Figure 3A, we present the performance of GPT-4 across 6 different types of hemorrhages, where red indicates the annotated hemorrhagic lesions identified by GPT-4. The overall identification completeness percentage for the 6 types of hemorrhages by GPT-4 reached 72.6% (SD 18.6%). Among these, GPT-4 demonstrated the highest identification completeness percentages for epidural and intraparenchymal hemorrhages, achieving 89.0% (SD 19.1%) and 86.9% (SD 17.7%), respectively. The identification completeness percentages for acute subdural hemorrhages (74.4%, SD 33.8%), subarachnoid hemorrhages (71.5%, SD 36.5%), and complex hemorrhages (76.4%, SD 32.9%) were also relatively high, whereas it was very low for chronic subdural hemorrhages (37.3%, SD 37.5%; Figure 3B). Nevertheless, the misidentification percentages for complex hemorrhages (54.0%, SD 28.0%), epidural hemorrhages (50.2%, SD 22.7%), and subarachnoid hemorrhages (50.5%, SD 29.2%) were high (>50%), whereas they were relatively low (<40%) for acute subdural hemorrhages (32.6%, SD 26.3%), chronic subdural hemorrhages (40.3%, SD 27.2%), and intraparenchymal hemorrhages (26.2%, SD 23.8%; Figure 3C). These results suggest that GPT-4 has a good capability to recognize some types of acute hemorrhages but lacks the ability to comprehensively identify chronic hemorrhages.

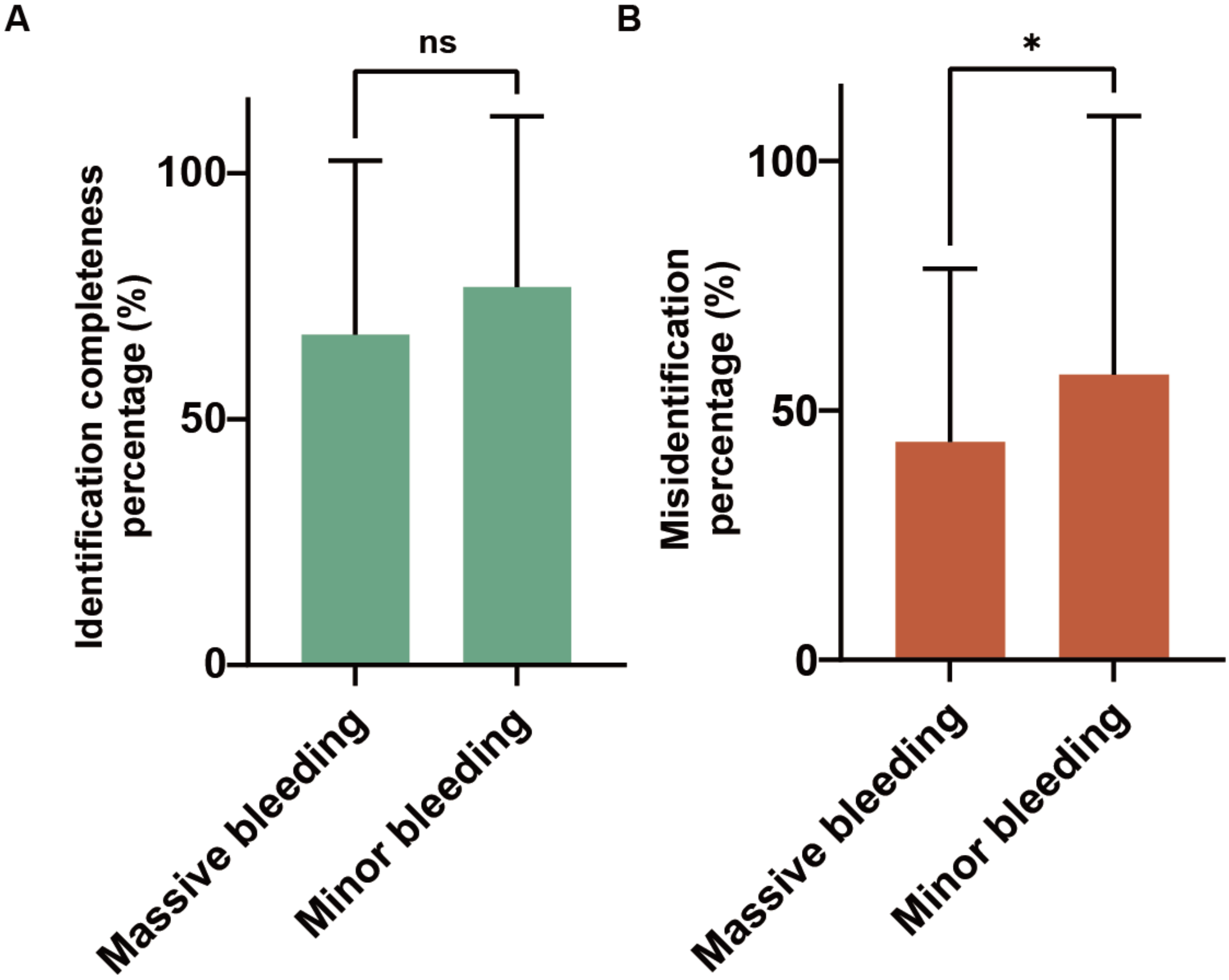
Further, we defined identification completeness percentages of >90%, from 50% to 90%, and ≤50% as high, moderate, and low identification, respectively. For epidural hemorrhages, the proportion of high identification reached 66% (21/32), and the proportion of moderate to high identification reached 91% (29/32), which further illustrated GPT-4’s strong capability in identifying epidural hemorrhages (Figure 3D). For intraparenchymal hemorrhages, the proportion of high identification was 48% (15/31), and the proportion of moderate to high identification was 94% (29/31), indicating that GPT-4 also had a good ability to recognize intraparenchymal hemorrhages (Figure 3D). In the cases of acute subdural hemorrhages and complex hemorrhages, approximately one-third of each set of images were categorized into high (12/42, 29% and 13/37, 37%, respectively), moderate (17/42, 40% and 14/37, 38%), and low (13/42, 31% and 10/37, 27%) identification; for subarachnoid hemorrhages, there was a substantial polarization (high identification at 21/36, 58% and low identification at 12/36, 33%), suggesting that GPT-4’s recognition of these 3 types of hemorrhages was not stable (Figure 3D). Finally, in chronic subdural hemorrhages, the proportion of low identification reached up to 73% (22/30), further showing GPT-4’s challenges in accurately identifying chronic subdural hemorrhages (Figure 3D).

Figure 3. Identification completeness and misidentification in 6 types of cerebral hemorrhages. (A) Representative images of 6 types of cerebral hemorrhages, separated into the massive and minor bleeding groups using a benchmark of 5 mL for the volume of bleeding. (B and C) Identification completeness and misidentification percentage of different types of the cerebral hemorrhages (mean and SD); *P<.05, **P<.01, ***P<.001, and ****P<.0001 by 1-way ANOVA with Tukey multiple comparison test. (D) The distribution of different identification completeness percentage groups in 6 types of cerebral hemorrhages.



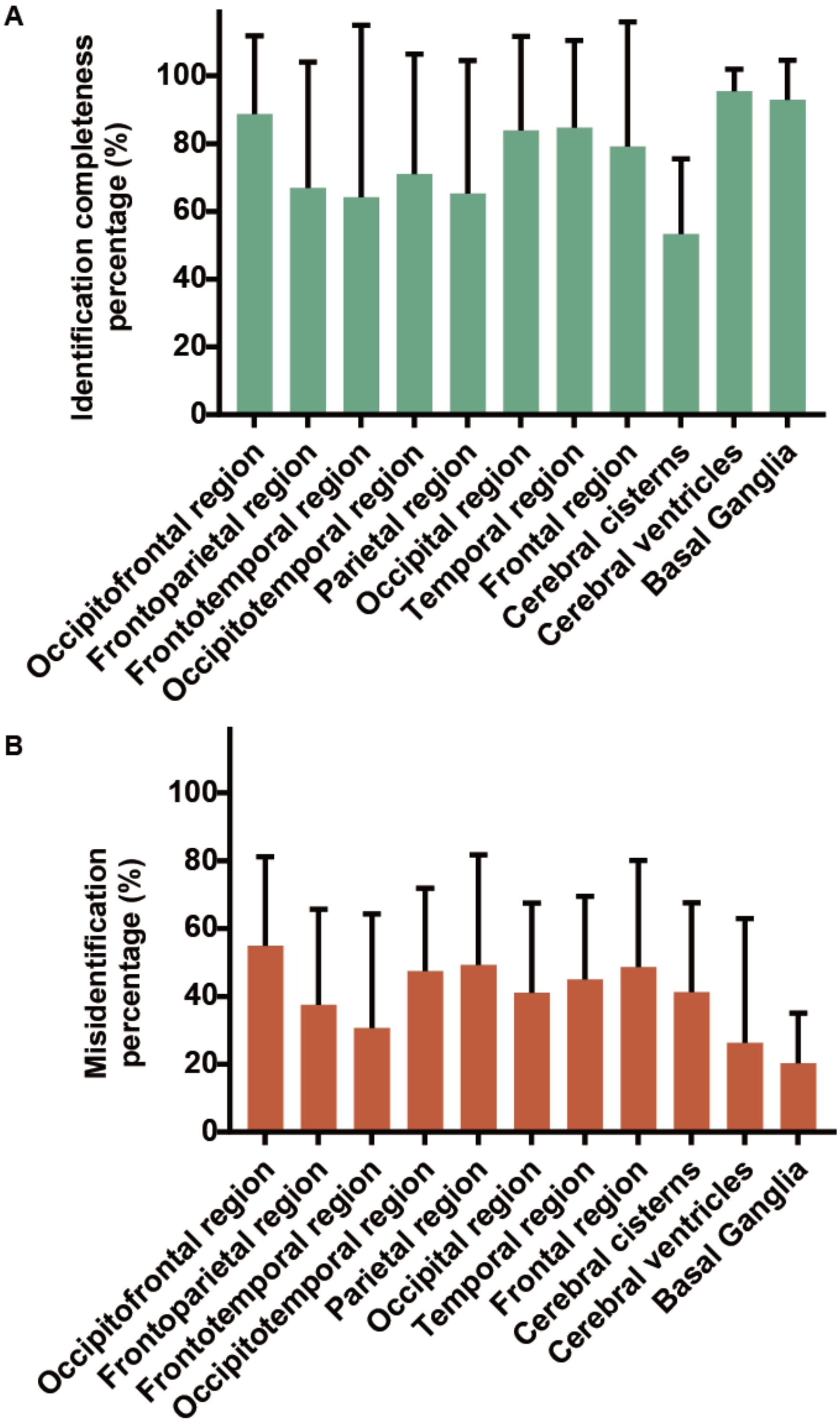
We classified cerebral hemorrhages into massive and minor bleeding, using a benchmark of 5 mL for the volume of bleeding. Surprisingly, there was no significant difference in the identification completeness percentages by GPT-4 between massive and minor bleeding (P=.06), indicating that the bleeding volume, within a certain range, did not affect GPT-4's ability to identify cerebral hemorrhages (Figure 4A). However, the misidentification percentage was higher for minor bleeding than for massive bleeding, suggesting that GPT-4 was more adept at recognizing substantial cerebral hemorrhages (P=.04; Figure 4B).

Figure 4. Identification completeness and misidentification percentage in the massive and minor bleeding groups. Cerebral hemorrhages were classified into massive and minor bleeding, using a benchmark of 5 mL for the volume of bleeding (mean and SD); *P<.05 and ns=P>.05 by 2-tailed t test. ns: no significance.



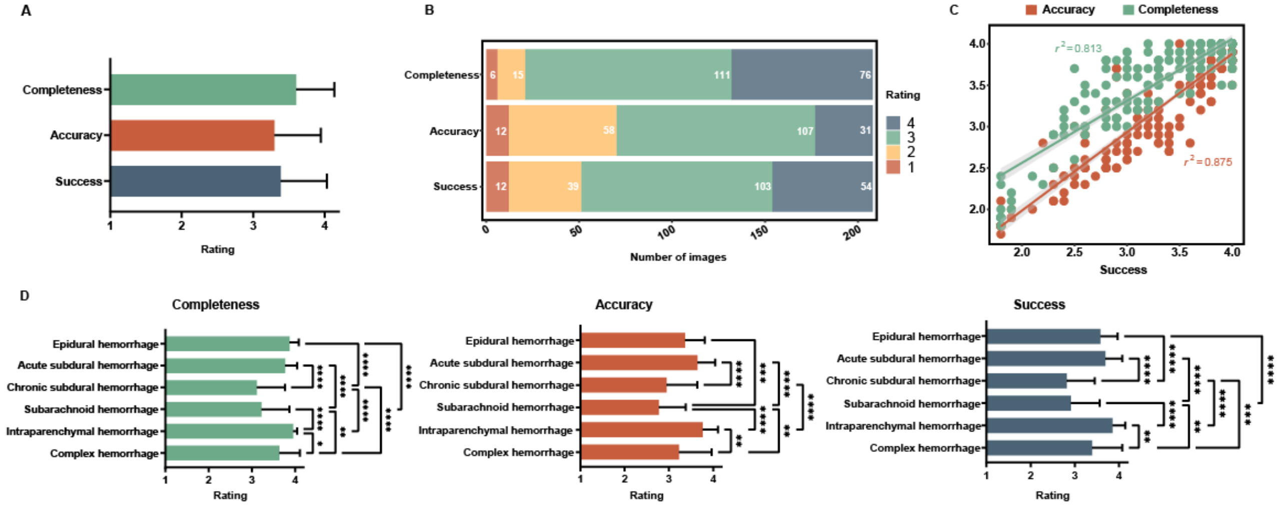
When we further performed analysis based on the location of hemorrhage within the brain, it was observed that GPT-4 demonstrated relatively higher identification completeness percentages and lower misidentification percentages in identifying hemorrhages in the cerebral ventricles and basal ganglia (Figure 5A and B). Conversely, the identification completeness percentage was comparatively lower for recognizing hemorrhages in the cerebral cisterns, and there was a higher misidentification percentage for hemorrhages in the occipitofrontal regions (Figure 5A and B). However, overall, there were no significant differences in the identification completeness percentage and misidentification percentage among different locations (all *P*>.05; Figure 5A and B). This indicated that GPT-4 did not show a marked preference for hemorrhages in specific areas but was relatively more proficient at identifying hemorrhages within the cerebral ventricles and basal ganglia.

Figure 5. Identification completeness and misidentification percentage in the different locations of hemorrhage within the brain. Cerebral hemorrhages were classified according to different hemorrhagic locations (mean and SD); by 1-way ANOVA with Tukey multiple comparison test.



Based on the aforementioned results, we examined how radiologists viewed the performance of GPT-4 in identifying cerebral hemorrhages. Overall, radiologists expressed relative acceptance in terms of identification completeness (3.60, SD 0.54), accuracy (3.30, SD 0.65), and success ((3.38, SD 0.64; Figure 6A). Specifically, 89.9% (187/208) of CT images were accepted in terms of identification completeness; 66.3% (138/208) of CT images were accepted in terms of identification accuracy; and 75.5% (157/208) of CT images were accepted in terms of identification success (Figure 6B). Moreover, at the same level of identification success, the average rating for identification completeness was higher than that for identification accuracy, indicating that radiologists placed more importance on GPT-4's ability to fully identify hemorrhagic lesions and relative tolerance for false positives (Figure 6C). Generally, whether it was in terms of identification completeness, accuracy, or success, the majority of CT scans were only deemed “relatively acceptable” (from 103/208, 49.5% to 111/208, 53.4%; Figure 6B). However, the proportion categorized as “completely unacceptable” was also very low (from 31/208, 14.9% to 76/208, 36.5%; Figure 6B). Further analysis was conducted based on different types of hemorrhages. The identification completeness, accuracy, and success of intraparenchymal hemorrhages and acute subdural hemorrhages received the highest acceptance from radiologists. Chronic subdural hemorrhages and subarachnoid hemorrhages had the lowest level of acceptance, with the identification completeness nearly reaching “relatively unacceptable” and both the identification accuracy and success being rated as “relatively unacceptable” (Figure 6D).

Figure 6. Evaluation on the CT identification by 10 radiologists. (A) The rating of the identification completeness, accuracy, and success (mean and SD). (B) The distribution of rating from identification completeness, accuracy and success. (C) The correlation between the identification completeness (green) or accuracy (red) and success. (D) The rating of identification completeness, accuracy, and success in 6 types of cerebral hemorrhages (mean and SD); *P<.05, **P<.01, ***P<.001, and ****P<.0001 by 1-way ANOVA with Tukey multiple comparison test. CT: computed tomography.



Discussion

In this study, we explored the possibility of applying GPT-4 to identify cerebral hemorrhages directly from cranial CT images, implying that NLP models like GPT-4 could be of clinical and commercial value in practice.

Overall, GPT-4 can identify most cerebral hemorrhage CT images after simple language-based training, achieving an overall identification completeness percentage of 72.6% (SD 18.6%). If chronic subdural hemorrhage is not considered, then the identification completeness percentage improves to 79.6% (SD 7.8%). This is significantly higher than our initial expectations prior to conducting this study, and radiologists have expressed an acceptable attitude toward this outcome (Figure 6A).

Specifically, in terms of different types of cerebral hemorrhages, GPT-4 demonstrates relatively high identification completeness percentages for epidural and intraparenchymal hemorrhages, with 89.0% (SD 19.1%) and 86.9% (SD 17.7%), respectively (Figure 3B). However, its identification completeness percentage for chronic subdural hemorrhage is very low, at only 37.3% (SD 37.5%; Figure 3B). The reason for this is that acute hematomas have a higher CT value compared to normal brain tissue, making the hematoma on CT images appear “whiter,” thus easier for GPT-4 to differentiate from normal brain tissue. In contrast, chronic hematomas have CT values closer to that of normal brain tissue, making it challenging for GPT-4 to identify them based on image brightness variations. Radiologists diagnose chronic subdural hematomas more on the basis of changes in brain tissue symmetry, hematoma shape, and the absence of brain sulci and gyri structures in the hematoma, rather than just changes in CT values [31]. These are aspects that the current version of GPT-4 struggles to comprehend. Additionally, this may also be related to the inadequacy of the prompts provided to GPT-4. We attempted to instruct GPT-4 to recognize changes in the symmetry of brain tissue structures, but this resulted in even poorer identification completeness for chronic subdural hemorrhage (data not included). The current version of GPT-4 may not yet fully grasp human anatomical structures and their alterations. This suggests that if NLP models are to be applied in clinical scenarios, they may need enhanced training in fundamental medical knowledge.

In terms of identification accuracy, GPT-4’s performance was not consistent. Compared to minor bleeding, GPT-4 was relatively more accurate in identifying massive bleeding (Figure 4B). This might be because it is challenging for GPT-4 to distinguish smaller hematomas from normal brain tissue due to their limited size. GPT-4 demonstrated relatively greater precision in recognizing hematomas in the cerebral ventricles and basal ganglia, possibly because these hematomas were located in the central part of the cranial CT scans, where their grayscale values were less likely to be negatively affected by the skull bones (Figure 5B). Therefore, for larger and relatively isolated hemorrhagic lesions, GPT-4 can identify them with more accuracy.

However, the current iteration of GPT-4 is not yet ready for clinical application and has the following limitations. First, the identification completeness and accuracy for some types of intracranial hemorrhages are relatively low. This directly prevents GPT-4 from being applied in clinical settings. Second, at present, GPT-4 can only recognize a single CT image plane and cannot recognize a series of continuous CT images. The specific choice of which image plane to use still requires a doctor's experience to decide. Third, a simple language-based training is needed before using GPT-4; therefore, inappropriate prompt words can greatly affect the recognition effectiveness of GPT-4. Fourth, the use of GPT-4 to identify intracranial hemorrhage also involves ethical and legal issues, which may not be universally accepted.

Nevertheless, we believe that the future trend will see AI assisting clinicians and enhancing efficiency in their work. Imagine scenarios like this: in hospitals lacking radiologists, assistants without medical background could upload patients' head CT images to GPT-4 for preliminary screening; young radiologists could use GPT-4 to analyze cerebral hemorrhage CT images, helping to avoid missed diagnoses; and in extraordinary situations like wars, earthquakes, or maritime disasters, even laypeople without medical training would not be perplexed when facing head CT scans. From the evaluation of 10 radiologists, it appears that they are more concerned about GPT-4's ability to identify hematomas, showing more tolerance toward misidentifications (Figure 6C). After all, for NLP models, not missing a hematoma holds greater clinical significance. In addition to clinical work, GPT-4 could potentially play a more significant role in medical education. In the traditional learning of CT or other radiologic imaging, medical students typically only receive images and the correct answers. Identifying the exact part of the image that represents the correct answer often leaves students confused. This approach focuses more on the result than the learning process. However, GPT-4 can assist in identifying abnormalities in images, and we can even envisage a future where GPT-4 could provide the correct diagnosis step by step. This shift would make medical education more process oriented and efficient.

In conclusion, this study represents the first exploration of applying GPT-4 to the identification of cerebral hemorrhages in cranial CT images, as well as the first attempt to utilize GPT-4 in the identification of radiological images. The current version of GPT-4 is capable of identifying a small portion of cerebral hemorrhages, indicating that it has a foundation for computer-assisted diagnosis. We believe that in the future, NLP models like GPT-4 will demonstrate immense clinical and commercial value.

Acknowledgments

This study was supported by grants from the Program of Shanghai Academic Research Leader (21XD1422400 to JF), Shanghai Municipal Education Commission-Gaofeng Clinical Medicine Support (02.101005.001.30.30A to JF), and the Project of Shanghai Medical And Health Development Foundation (20224Z0012 to JF).

We extend our heartfelt gratitude to the other 8 dedicated radiologists from the Department of Radiology at Ren Ji Hospital, Shanghai Jiao Tong University School of Medicine, and Minhang Hospital, Fudan University. Their invaluable participation in evaluation significantly contributed to the depth and breadth of this study.

Authors' Contributions

Conceptualization – WW, JF
Data curation – DZ, ZM, RG
Formal analysis – DZ, ZM, RG, ZH
Funding acquisition – JF
Investigation – DZ, LL, YL
Methodology – DZ, ZM, RG, LL, YL, J Huang, WW, JF
Supervision – WW, JF
Validation – ZM, RG, YH, J Hui
Writing – original draft – DZ, JJ, WW, JF
Writing – review & editing – JJ

Conflicts of Interest

None declared.

Multimedia Appendix 1

Representative steps of computed tomography (CT) images inputted into GPT-4. (A) The initial natural language prompt is inputted with the CT image into GPT-4, assisting GPT-4 to inspect the original picture and give the annotation back. Users' feedbacks help GPT-4 successfully annotate the bleeding area precisely. (B) The training CT image.
[PNG File , 638 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Criteria and categories used for rating of annotated images by radiologists.
[DOCX File , 24 KB-Multimedia Appendix 2]

References

1. Stocchetti N, Zanier ER. Chronic impact of traumatic brain injury on outcome and quality of life: a narrative review. Crit Care. Jun 21, 2016;20(1):148. [FREE Full text] [doi: 10.1186/s13054-016-1318-1] [Medline: 27323708]
2. Bayen E, Jourdan C, Ghout I, Darnoux E, Azerad S, Vallat-Azouvi C, et al. Objective and subjective burden of informal caregivers 4 years after a severe traumatic brain injury: results from the PariS-TBI study. J Head Trauma Rehabil. 2016;31(5):E59-E67. [doi: 10.1097/HTR.0000000000000079] [Medline: 24992640]
3. Humphreys I, Wood RL, Phillips CJ, Macey S. The costs of traumatic brain injury: a literature review. Clinicoecon Outcomes Res. Jun 26, 2013;5:281-287. [FREE Full text] [doi: 10.2147/CEOR.S44625] [Medline: 23836998]
4. Lee WJ, Shah Y, Ku A, Patel N, Salvador M. Evaluating health disparities in radiology practices in New Jersey: exploring radiologist geographical distribution. Cureus. Aug 2023;15(8):e43474. [FREE Full text] [doi: 10.7759/cureus.43474] [Medline: 37583547]
5. Zhang J, Han X, Yang Z, Wang Z, Zheng J, Yang Z, et al. Radiology residency training in China: results from the first retrospective nationwide survey. Insights Imaging. Feb 17, 2021;12(1):25. [FREE Full text] [doi: 10.1186/s13244-021-00970-2] [Medline: 33595737]
6. Bruls RJM, Kwee RM. Workload for radiologists during on-call hours: dramatic increase in the past 15 years. Insights Imaging. Nov 23, 2020;11(1):121. [FREE Full text] [doi: 10.1186/s13244-020-00925-z] [Medline: 33226490]
7. Wang YE, Liu M, Jin L, Lungren MP, Grimm LJ, Zhang Z, et al. Radiology education in China. J Am Coll Radiol. Mar 2013;10(3):213-219. [doi: 10.1016/j.jacr.2012.11.006] [Medline: 23571062]
8. Mohsen H, El-Dahshan ESA, El-Horbaty ESM, Salem ABM. Classification using deep learning neural networks for brain tumors. Future Computing and Informatics Journal. Jun 2018;3(1):68-71. [doi: 10.1016/j.fcij.2017.12.001]
9. Fernando T, Gammulle H, Denman S, Sridharan S, Fookes C. Deep learning for medical anomaly detection – a survey. ACM Comput Surv. Jul 18, 2021;54(7):1-37. [doi: 10.1145/3464423]
10. Hesamian MH, Jia W, He X, Kennedy P. Deep learning techniques for medical image segmentation: achievements and challenges. J Digit Imaging. Aug 29, 2019;32(4):582-596. [FREE Full text] [doi: 10.1007/s10278-019-00227-x] [Medline: 31144149]
11. Guo J, Li B. The application of medical artificial intelligence technology in rural areas of developing countries. Health Equity. Aug 2018;2(1):174-181. [FREE Full text] [doi: 10.1089/heq.2018.0037] [Medline: 30283865]
12. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: systematic review. JMIR Med Educ. Jun 30, 2020;6(1):e19285. [FREE Full text] [doi: 10.2196/19285] [Medline: 32602844]
13. Zhang W, Cai M, Lee HJ, Evans R, Zhu C, Ming C. AI in medical education: global situation, effects and challenges. Educ Inf Technol. Jul 10, 2023;29(4):4611-4633. [doi: 10.1007/s10639-023-12009-8]
14. Duong MT, Rauschecker AM, Rudie JD, Chen P, Cook TS, Bryan RN, et al. Artificial intelligence for precision education in radiology. Br J Radiol. Nov 2019;92(1103):20190389. [FREE Full text] [doi: 10.1259/bjr.20190389] [Medline: 31322909]
15. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. 2016. Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016:779-788; Las Vegas, NV. [doi: 10.1109/cvpr.2016.91]
16. Ertuğrul Ö, Akıl MF. Detecting hemorrhage types and bounding box of hemorrhage by deep learning. Biomedical Signal Processing and Control. Jan 2022;71:103085. [doi: 10.1016/j.bspc.2021.103085]
17. Lee H, Huang C, Yune S, Tajmir SH, Kim M, Do S. Machine friendly machine learning: interpretation of computed tomography without image reconstruction. Sci Rep. Oct 29, 2019;9(1):15540. [FREE Full text] [doi: 10.1038/s41598-019-51779-5] [Medline: 31664075]
18. Introducing ChatGPT. OpenAI. URL: https://openai.com/blog/chatgpt [accessed 2024-06-10]
19. GPT-4. OpenAI. URL: https://openai.com/gpt-4 [accessed 2024-06-10]
20. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. ACM Comput Surv. Mar 03, 2023;55(12):1-38. [doi: 10.1145/3571730]
21. Chelli M, Descamps J, Lavoué V, Trojani C, Azar M, Deckert M, et al. Hallucination rates and reference accuracy of ChatGPT and Bard for systematic reviews: comparative analysis. J Med Internet Res. May 22, 2024;26:e53164. [FREE Full text] [doi: 10.2196/53164] [Medline: 38776130]
22. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. Radiology. May 01, 2023;307(4):e230725. [doi: 10.1148/radiol.230725] [Medline: 37014240]
23. Fink MA, Bischoff A, Fink CA, Moll M, Kroschke J, Dulz L, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer. Radiology. Sep 2023;308(3):e231362. [doi: 10.1148/radiol.231362] [Medline: 37724963]
24. Santana LDM, Floresta L, Alves, dos Santos MAL, Barbosa BF, de Vasconcellos S, et al. Can GPT-4 be a viable alternative for discussing complex cases in digital oral radiology? a critical analysis. EXCLI J. Aug 1, 2023;22:749-751. [FREE Full text] [doi: 10.17179/excli2023-6373] [Medline: 37662708]

25. 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. Oct 2023;20(10):990-997. [FREE Full text] [doi: 10.1016/j.jacr.2023.05.003] [Medline: 37356806]
26. Truhn D, Weber CD, Braun BJ, Bressem K, Kather JN, Kuhl C, et al. A pilot study on the efficacy of GPT-4 in providing orthopedic treatment recommendations from MRI reports. Sci Rep. Nov 17, 2023;13(1):20159. [FREE Full text] [doi: 10.1038/s41598-023-47500-2] [Medline: 37978240]
27. Horiuchi D, Tatekawa H, Shimono T, Walston SL, Takita H, Matsushita S, et al. Accuracy of ChatGPT generated diagnosis from patient's medical history and imaging findings in neuroradiology cases. Neuroradiology. Jan 2024;66(1):73-79. [doi: 10.1007/s00234-023-03252-4] [Medline: 37994939]
28. Tang XN, Berman AE, Swanson RA, Yenari MA. Digitally quantifying cerebral hemorrhage using Photoshop and Image J. J Neurosci Methods. Jul 15, 2010;190(2):240-243. [FREE Full text] [doi: 10.1016/j.jneumeth.2010.05.004] [Medline: 20452374]
29. Kothari RU, Brott T, Broderick JP, Barsan WG, Sauerbeck LR, Zuccarello M, et al. The ABCs of measuring intracerebral hemorrhage volumes. Stroke. Aug 1996;27(8):1304-1305. [doi: 10.1161/01.str.27.8.1304] [Medline: 8711791]
30. Webb AJ, Ullman NL, Morgan TC, Muschelli J, Kornbluth J, Awad IA, et al. MISTIECLEAR Investigators. Accuracy of the ABC/2 score for intracerebral hemorrhage: systematic review and analysis of MISTIE, CLEAR-IVH, and CLEAR III. Stroke. Sep 2015;46(9):2470-2476. [FREE Full text] [doi: 10.1161/STROKEAHA.114.007343] [Medline: 26243227]
31. Yadav Y, Parihar V, Namdev H, Bajaj J. Chronic subdural hematoma. Asian J Neurosurg. Sep 20, 2016;11(4):330-342. [FREE Full text] [doi: 10.4103/1793-5482.145102] [Medline: 27695533]

Abbreviations

AI: artificial intelligence
CT: computed tomography
NLP: natural language processing

Edited by G Eysenbach; submitted 23.03.24; peer-reviewed by W Wu, J Mistry, A Wani, D Ghosh; comments to author 24.04.24; revised version received 19.06.24; accepted 31.08.24; published 26.09.24

Please cite as:
Zhang D, Ma Z, Gong R, Lian L, Li Y, He Z, Han Y, Hui J, Huang J, Jiang J, Weng W, Feng J
Using Natural Language Processing (GPT-4) for Computed Tomography Image Analysis of Cerebral Hemorrhages in Radiology: Retrospective Analysis
J Med Internet Res 2024;26:e58741
URL: https://www.jmir.org/2024/1/e58741
doi: 10.2196/58741
PMID: