SIIM CMIMI24
Oct 21-22 | Boston, MA
Conference On Machine Intelligence In Medical Imaging

Regression in GPT-4 Turbo's Diagnostic Accuracy for Generating Radiology Differential Diagnoses

David Li, MD, London Health Sciences Center; Kartik Gupta, MSc; Mousumi Bhaduri, MBBS, DMRD, DNB, DABR; Paul Sathiadoss, MBBS; Sahir Bhatnagar, PhD; Jaron Chong, MD, MHI

Introduction/Background

Large language models (LLMs) have demonstrated impressive capabilities across a variety of domains; however, their effectiveness in clinical tasks, such as generating differential diagnoses, remains underexplored. This study evaluates the diagnostic accuracy of GPT-4 Turbo, an advanced generative pre-trained transformer (GPT), in analyzing Radiology Diagnosis Please cases. These cases encompass a broad range of pathologies, reflecting the complexities of diagnostic radiology. We hypothesize that GPT-4 Turbo will outperform its predecessors in generating accurate differential diagnoses.

Methods/Intervention

This study was exempt from institutional review board review due to the use of publicly available data. We retrospectively compiled a test set of 287 Radiology Diagnosis Please cases from August 1998 to July 2023, excluding cases with information leaks. Patient histories, imaging findings, and ground truth diagnoses were extracted. The latest version of GPT-4 Turbo (April 2024 release) was evaluated. Diagnostic accuracy was assessed by generating the top five differential diagnoses based on text inputs of history, imaging findings, and their combination. A panel of three radiologists, averaging 13 years of experience, evaluated blinded differentials and resolved discrepancies through mediated discussion.

Results/Outcome

GPT-4 Turbo's diagnostic accuracy based on the history, imaging findings, and both combined were 43/287 (15%), 119/287 (41%), and 132/287 (46%), respectively (Table 1). Accuracy varied across subspecialties, ranging from 0/26 (0%) in genitourinary cases to 4/6 (67%) in obstetrics cases. Qualitative observations of diagnostic regression included lower rankings of correct diagnoses and the omission of eponyms and previously accurate diagnoses (Fig. 1).

Conclusion

This clinical validation study identifies an unexpected regression in the diagnostic accuracy of GPT-4 Turbo compared to previously published benchmarks for GPT-4 and GPT-3.5. These results highlight the need for additional fine-tuning to enhance GPT-4 Turbo's performance and ensure its effectiveness before clinical deployment.

Statement of Impact

This clinical validation study underscores the importance of exercising caution when integrating LLMs into diagnostic workflows. The regression in GPT-4 Turbo's performance suggests that foundational models require additional fine-tuning with medical datasets. Rigorous validation of LLMs is crucial to establish their effectiveness and reliability before widespread clinical adoption. With continuous improvements, LLMs have the potential to become valuable decision support tools for radiologists.

Upload Supplemental Documents

Subspecialty	History	Imaging Findings	History and Imaging Findings
Total	38/287 (13%)	109/287 (38%)	120/287 (41%)
Breast	1/10 (10%)	2/10 (20%)	4/10 (40%)
Cardiovascular	2/17 (12%)	10/17 (59%)	11/17 (65%)
Chest	5/35 (14%)	11/35 (31%)	14/35 (40%)
Gastrointestinal	5/56 (8.9%)	20/56 (36%)	19/56 (34%)
Genitourinary	0/26 (0%)	12/26 (46%)	10/26 (38%)
Head and neck	2/9 (22%)	4/9 (44%)	4/9 (44%)
Musculoskeletal	3/30 (10%)	8/30 (27%)	9/30 (30%)
Neuroradiology	12/46 (26%)	11/46 (24%)	17/46 (37%)
Obstetric	1/6 (17%)	4/6 (67%)	3/6 (50%)
Pediatric	7/52 (13%)	27/52 (52%)	29/52 (56%)

Table 1: Diagnostic accuracy of GPT-4 Turbo (April 2024) overall and by subspecialty for 287 Radiology Diagnosis Please cases.

Case	GPT-4 March 2023 Differential Diagnoses	GPT-4 Turbo April 2024 Differential Diagnoses
#142: Susac syndrome *Performance regression:* *- Three previously correct diagnoses omitted* *- Eponym no longer proposed*	**History only:** ['Optic neuritis', 'Migraine with aura', 'Vestibular neuritis', 'Multiple sclerosis', 'Susac syndrome'] **Imaging findings only:** ['Marchiafava-Bignami disease', 'Multiple sclerosis', 'Susac syndrome', 'Small vessel ischemic disease', 'Progressive multifocal leukoencephalopathy'] **History and imaging findings:** 'Multiple sclerosis', 'Neuromyelitis optica', 'Susac syndrome', 'Acute disseminated encephalomyelitis', 'Central nervous system vasculitis'	**History only:** ['multiple sclerosis', 'vestibular schwannoma', 'migraine with brainstem aura', 'stroke', 'neuromyelitis optica spectrum disorder'] **Imaging findings only:** ['Multiple sclerosis', 'Marchiafava-Bignami disease', 'Lymphoma', 'Progressive multifocal leukoencephalopathy', 'Acute disseminated encephalomyelitis'] **History and imaging findings:** ['multiple sclerosis', 'acute disseminated encephalomyelitis', 'neuromyelitis optica spectrum disorder', 'central nervous system vasculitis', 'Lyme disease'] **Correct Diagnosis Not Proposed**
#144: Dysplastic cerebellar gangliocytoma (Lhermitte-Duclos disease) *Performance regression:* *- Two previously correct diagnoses and one non-specific diagnosis omitted*	**History only:** ['*cerebellar tumor*', 'pseudotumor cerebri', 'Chiari malformation', 'cerebellar abscess', 'cerebellar infarction'] **Imaging findings only:** ['Lhermitte-Duclos disease', 'medulloblastoma', 'ependymoma', 'astrocytoma', 'hemangioblastoma'] **History and imaging findings:** ['Lhermitte-Duclos disease', 'medulloblastoma', 'ependymoma', 'astrocytoma', 'hemangioblastoma']	**History only:** ['posterior fossa tumor', 'Chiari malformation', 'hydrocephalus', 'cerebellar degeneration', 'meningitis'] **Imaging findings only:** ['medulloblastoma', 'ependymoma', 'astrocytoma', 'hemangioblastoma', 'cerebellar metastasis'] **History and imaging findings:** ['medulloblastoma', 'ependymoma', 'astrocytoma', 'hemangioblastoma', 'cerebellar metastasis'] **Correct Diagnosis Not Proposed**
#279: Central-variant posterior reversible encephalopathy syndrome *Performance regression:* *- One previously correct diagnosis omitted* *- Lower rank position of the correct diagnosis in the differential*	**History only:** ['Posterior reversible encephalopathy syndrome', 'Cerebral lupus vasculitis', 'Seizure-related injury', 'Hypertensive encephalopathy', 'Ischemic stroke'] **Imaging findings only:** ['Posterior reversible encephalopathy syndrome', 'Acute disseminated encephalomyelitis', 'Central pontine myelinolysis', 'Bickerstaff brainstem encephalitis', 'Wernicke encephalopathy'] **History and imaging findings:** ['Posterior reversible encephalopathy syndrome', 'Cerebral lupus vasculitis', 'Central pontine myelinolysis', 'Acute disseminated encephalomyelitis', 'Hypertensive encephalopathy']	**History only:** ['neuropsychiatric systemic lupus erythematosus', 'central pontine myelinolysis', 'posterior reversible encephalopathy syndrome', 'acute disseminated encephalomyelitis', 'Wernicke encephalopathy'] **Imaging findings only:** ['Top of the basilar syndrome', 'Central pontine myelinolysis', 'Hypertensive encephalopathy', 'Cerebral fat embolism', 'Acute disseminated encephalomyelitis'] **History and imaging findings:** ['Central pontine myelinolysis', 'Posterior reversible encephalopathy syndrome (PRES)', 'Acute disseminated encephalomyelitis (ADEM)', 'Cerebral venous sinus thrombosis', 'Lupus cerebritis']

Fig. 1: Examples illustrating the regression in diagnostic accuracy of GPT-4 Turbo (April 2024) compared to GPT-4 (March 2023).

Keywords

Large language model; Generative pre-trained transformer; GPT-4 Turbo; Clinical validation; Diagnostic accuracy; Differential diagnosis