Radiology

ORIGINAL RESEARCH • COMPUTER APPLICATIONS

Comparing GPT-3.5 and GPT-4 Accuracy and Drift in *Radiology* Diagnosis Please Cases

David Li, MD • Kartik Gupta, BMSc • Mousumi Bhaduri, MBBS, DMRD, DABR • Paul Sathiadoss, MBBS • Sahir Bhatnagar, PhD • Jaron Chong, MD, MHI

From the Department of Medical Imaging, London Health Sciences Centre, 800 Commissioners Rd E, London, ON, Canada N6A 5A5 (D.L., M.B., P.S., J.C.); Department of Medical Imaging, Schulich School of Medicine & Dentistry, Western University, London, Ontario, Canada (K.G., J.C.); and Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Quebec, Canada (S.B.). Received September 8, 2023; revision requested October 6; revision received November 24; accepted December 4. **Address correspondence to** J.C. (email: *Jaron.Chong@lhsc.on.ca*).

Conflicts of interest are listed at the end of this article.

Radiology 2024; 310(1):e232411 • https://doi.org/10.1148/radiol.232411 • Content code: AI • © RSNA, 2024

Large language models (LLMs), such as generative pretrained transformers (GPTs), have garnered attention in the past year due to their remarkable capacity to comprehend and generate human-like text, with perhaps the most well-known being ChatGPT (1). However, it remains unquantified to what extent advancements in successive GPT generations translate into enhanced diagnostic accuracy for radiology cases. This investigation aims to evaluate the diagnostic accuracy of GPT-3.5 and GPT-4 (OpenAI) in solving text-based *Radiology* Diagnosis Please cases. GPT-4 is the successor to GPT-3.5 and has demonstrated substantial improvements on numerous academic examinations (2).

Materials and Methods

This study adheres to the Checklist for Artificial Intelligence in Medical Imaging and was exempt from institutional review board review due to the use of public data (3). A retrospective analysis of *Radiology* Diagnosis Please cases from August 1998 to July 2023 was performed. The clinical history, imaging findings, and ground truth diagnosis were extracted. Cases disclosing the diagnosis were excluded. Diagnostic accuracy of the March and June 2023 snapshots (ie, a specific model version from a point in time) of GPT-3.5 (4) and GPT-4 (5) were assessed using the top five differential diagnoses generated from text inputs of history, findings, and both combined, with imaging findings originally characterized by radiologists. Default hyperparameters were applied, except for a temperature of 0 to maximize determinism. Three radiologists (J.C., P.S., and M.B., with 8, 8, and 23 years of experience, respectively) evaluated generated differentials, with discrepancies resolved by means of mediated discussion. A generalized estimating equation linear probability model with an exchangeable correlation structure was fit to estimate the time-dependent effects and 95% CIs of snapshot version on diagnostic accuracy, with adjustment for subspecialty.

Results

Of 315 cases, 28 were excluded due to disclosed diagnoses for a final sample of 287 cases. Overall, GPT-4's accuracy improved significantly compared with GPT-3.5 by 19.8 percentage points (95% CI: 15, 25) in March and 11.1 percentage points (95% CI: 6, 17) in June (Tables 1, 2). Within models, for GPT-4, from March to June, there was a statistically significant decrease in accuracy (accuracy, −5.92 percentage points [95% CI: −10, −2]). For GPT-3.5, from March to June, there was an increase in accuracy that was not statistically significant (accuracy, +2.79 percentage points [95% CI: −1, 6]). Of the

Table 1: Overall and Per-Subspecialty Diagnostic Accuracy of GPT-3.5 and GPT-4 June 2023 Snapshots on 287 *Radiology* Diagnosis Please Cases

Model and Subspecialty	Clinical History Only	Imaging Findings Only	History and Findings
GPT-3.5	31/287 (10.8)	102/287 (35.5)	115/287 (40.1)
Breast	2/10 (20)	4/10 (40)	3/10 (30)
Cardiovascular	1/17 (5.9)	11/17 (65)	12/17 (71)
Chest	3/35 (8.6)	13/35 (37)	14/35 (40)
Gastrointestinal	3/56 (5.4)	17/56 (30)	20/56 (36)
Genitourinary	0/26 (0)	9/26 (35)	11/26 (42)
Head and neck	1/9 (11)	6/9 (67)	7/9 (78)
Musculoskeletal	2/30 (6.7)	8/30 (27)	8/30 (27)
Neuroradiology	10/46 (22)	14/46 (30)	17/46 (37)
Obstetric	0/6 (0)	0/6 (0)	1/6 (17)
Pediatric	9/52 (17)	20/52 (38)	22/52 (42)
GPT-4	49/287 (17.1)	133/287 (46.3)	147/287 (51.2)
Breast	2/10 (20)	4/10 (40)	4/10 (40)
Cardiovascular	1/17 (5.9)	13/17 (76)	12/17 (71)
Chest	6/35 (17)	14/35 (40)	16/35 (46)
Gastrointestinal	3/56 (5.4)	26/56 (46)	25/56 (45)
Genitourinary	0/26 (0)	7/26 (27)	8/26 (31)
Head and neck	3/9 (33)	7/9 (78)	8/9 (89)
Musculoskeletal	5/30 (17)	13/30 (43)	12/30 (40)
Neuroradiology	16/46 (35)	18/46 (39)	24/46 (52)
Obstetric	1/6 (17)	3/6 (50)	4/6 (67)
Pediatric	12/52 (23)	28/52 (54)	34/52 (65)

Note.—Data are numbers of cases, with percentages in parentheses. Each case was categorized into body systems after review of the original case images and diagnosis. In multisystem cases, the initiating image body system was selected.

Table 2: Overall and Per-Subspecialty Diagnostic Accuracy of GPT-3.5 and GPT-4 March 2023 Snapshots on 287 *Radiology* Diagnosis Please Cases

Model and Subspecialty	Clinical History Only	Imaging Findings Only	History and Findings
GPT-3.5	29/287 (10.1)	101/287 (35.2)	107/287 (37.3)
Breast	2/10 (20)	2/10 (20)	3/10 (30)
Cardiovascular	1/17 (5.9)	10/17 (59)	11/17 (65)
Chest	2/35 (5.7)	14/35 (40)	14/35 (40)
Gastrointestinal	3/56 (5.4)	18/56 (32)	19/56 (34)
Genitourinary	0/26 (0)	8/26 (31)	6/26 (23)
Head and neck	1/9 (11)	4/9 (44)	5/9 (56)
Musculoskeletal	3/30 (10)	8/30 (27)	9/30 (30)
Neuroradiology	9/46 (20)	14/46 (30)	16/46 (35)
Obstetric	0/6 (0)	2/6 (33)	2/6 (33)
Pediatric	8/52 (15)	21/52 (40)	22/52 (42)
GPT-4	42/287 (14.6)	132/287 (46.0)	164/287 (57.1)
Breast	2/10 (20)	4/10 (40)	4/10 (40)
Cardiovascular	1/17 (5.9)	12/17 (71)	11/17 (65)
Chest	4/35 (11)	12/35 (34)	19/35 (54)
Gastrointestinal	3/56 (5.4)	26/56 (46)	27/56 (48)
Genitourinary	2/26 (7.7)	12/26 (46)	15/26 (58)
Head and neck	2/9 (22)	5/9 (56)	8/9 (89)
Musculoskeletal	4/30 (13)	13/30 (43)	14/30 (47)
Neuroradiology	13/46 (28)	18/46 (39)	28/46 (61)
Obstetric	1/6 (17)	4/6 (67)	5/6 (83)
Pediatric	10/52 (19)	26/52 (50)	33/52 (63)

Note.—Data are numbers of cases, with percentages in parentheses. Each case was categorized into body systems after review of the original case images and diagnosis. In multisystem cases, the initiating image body system was selected.

10 subspecialties, with breast imaging as reference, the only subspecialty significantly associated with greater accuracy was head and neck cases (generalized estimating equation estimate: 0.428 [95% CI: 0.10, 0.76]). Across all subspecialties and snapshots, the average increase in diagnostic accuracy from GPT-3.5 to GPT-4 was +17.3% (SD, 15.6%; minimum, −11.5%; maximum, +50.0%) (Figure).

Discussion

Diagnosis Please cases could serve as a test for gauging performance drift, or changes in model performance over time, as they simulate complex, challenging, real-world clinical scenarios (6). Our results suggest performance drift between the March and June snapshots of GPT-3.5 and GPT-4. The overall increase in diagnostic accuracy between GPT-3.5 and GPT-4 moderately parallels that seen in other academic and professional examinations (2). If future LLMs exhibit similar performance increases, we anticipate that accuracy on Diagnosis Please cases may continue to increase, even without radiology-specific fine-tuning.

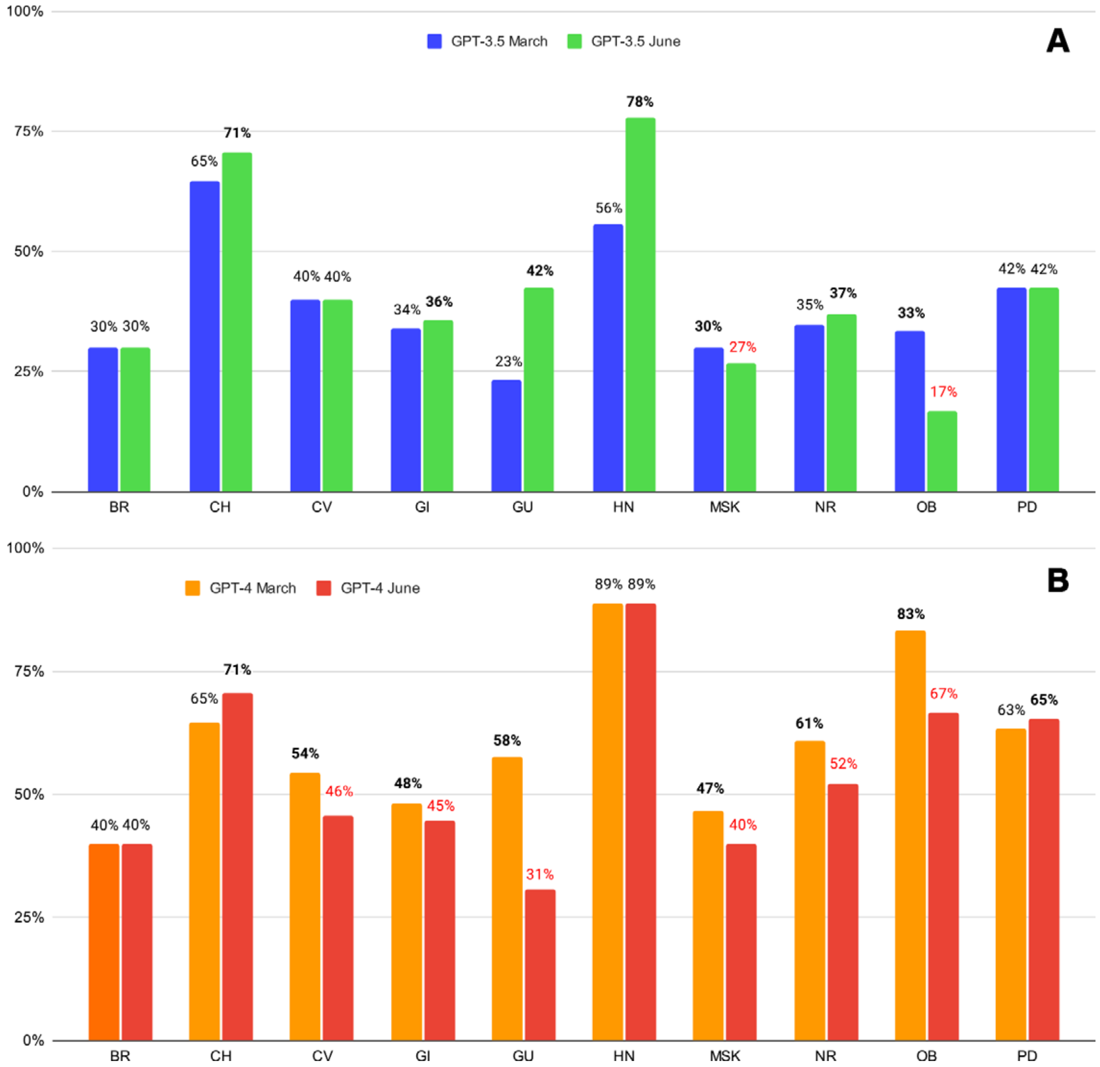
Our investigation demonstrated unexpected findings, notably that there was a statistically significant decrease in the diagnostic accuracy of the GPT-4 June snapshot. This observation echoes similar reports of GPT-4's performance varying between snapshots (7). This variability could stem from optimization on competing metrics, such as safety or inference speed, potentially leading to instability in real-world performance. Despite differences between this experimental setting and clinical practice, LLMs could potentially serve as a decision support tool in future diagnostic workflows, particularly for creatively broadening differential diagnoses under supervision by radiologists. Our study highlights the pressing need for more robust and continuous LLM monitoring systems before clinical deployment.

Author contributions: Guarantors of integrity of entire study, **K.G., J.C.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **D.L., K.G.**; clinical studies, **K.G.**; experimental studies, **D.L., K.G., M.B., P.S., J.C.**; statistical analysis, **D.L., K.G., S.B.**; and manuscript editing, **D.L., K.G., J.C.**

Disclosures of conflicts of interest: **D.L.** No relevant relationships. **K.G.** No relevant relationships. **M.B.** No relevant relationships. **P.S.** No relevant relationships. **S.B.** No relevant relationships. **J.C.** Member of the Health Canada Scientific Advisory Committee for Digital Health Technologies; chair of the Canadian Association of Radiologists AI Standing Committee.

References

1. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 2023;307(5):e230582.



Comparison stacked bar charts of diagnostic accuracy between March and June 2023 snapshots of GPT-3.5 and GPT-4 on 287 *Radiology* Diagnosis Please cases using text-based clinical history and findings. **(A)** For GPT-3.5, the diagnostic accuracy increased in five of 10 subspecialties, remained unchanged in three subspecialties, and decreased in two subspecialties between the March and June 2023 snapshots. **(B)** For GPT-4, the diagnostic accuracy increased in two of 10 subspecialties, remained unchanged in two subspecialties, and decreased in six subspecialties between the March and June 2023 snapshots. BR = breast, CH = chest, CV = cardiovascular, GI = gastrointestinal, GU = genitourinary, HN = head and neck, MSK = musculoskeletal, NR = neuroradiology, OB = obstetric, PD = pediatric.

2. OpenAI. GPT-4 Technical Report. arXiv 2303.08774 [preprint] https://arxiv.org/abs/2303.08774. Posted March 15, 2023. Updated March 27, 2023. Accessed September 2023.
3. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020;2(2):e200029.
4. GPT-3.5. OpenAI. https://platform.openai.com/docs/models/gpt-3-5. Accessed August 12, 2023.
5. GPT-4. OpenAI. https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo. Accessed August 12, 2023.
4. Ueda D, Mitsuyama Y, Takita H, et al. ChatGPT's diagnostic performance from patient history and imaging findings on the Diagnosis Please quizzes. Radiology 2023;308(1):e231040.
5. Chen L, Zaharia M, Zou J. How is ChatGPT's behavior changing over time? arXiv 2307.09009 [preprint] https://arxiv.org/abs/2307.09009. Posted July 18, 2023. Updated October 31, 2023. Accessed September 2023.