HealthManagement
Promoting Management and Leadership

Performance of ChatGPT in Radiology: Reliability, Repeatability, and Robustness in Question

OpenAI's ChatGPT, powered by generative pretrained transformer language models, has shown promise in various fields, including medicine and radiology, despite its default version not being fine-tuned for specific domains. Studies have demonstrated its potential in assisting with decision-making, protocol creation, and patient inquiries. However, a significant limitation is its tendency to generate incorrect responses, known as hallucinations, which affects its accuracy, particularly in medical contexts. While the stochastic nature of these models enhances response diversity and adaptability, it raises concerns regarding reliability, repeatability, and robustness, especially in radiology, where accuracy is crucial. Moreover, ChatGPT may exhibit overconfidence in its responses, posing risks, especially for novice users. Currently, it lacks a mechanism to convey its confidence level. Thus, a study recently published in Radiology was conducted to evaluate the reliability, repeatability, robustness, and confidence of GPT-3.5 and GPT-4 through repeated prompting with a radiology board–style examination.

Assessing the Performance of ChatGPT in Radiology: Reliability, Repeatability, and Robustness

In an exploratory prospective study, the default versions of ChatGPT (GPT-3.5 and GPT-4) were subjected to 150 radiology board–style multiple-choice text-based questions across three attempts, spaced by intervals of ≥1 month and then 1 week. The aim was to evaluate the reliability (accuracy over time) and repeatability (agreement over time) by comparing accuracy and answer choices between attempts. Additionally, the robustness (ability to withstand adversarial prompting) was tested by challenging ChatGPT three times with an adversarial prompt on the third attempt. Confidence ratings from 1–10 were collected after each challenge prompt and on the third attempt. The examination questions were standardised, matching those used in previous benchmarking studies, covering various radiology topics. The questions were classified based on Bloom's Taxonomy into lower-order and higher-order thinking types. This study expanded on previous research by assessing reliability, repeatability, robustness, and the relationship between confidence and accuracy over multiple attempts.

No Parameter Adjustment but Adversarial Prompting

The default versions of ChatGPT (GPT-3.5 and GPT-4) were utilized without parameter adjustment or prompt engineering in a study conducted between March 2023 and January 2024. The study involved entering each of the 150 radiology board–style questions, along with their four answer choices, into ChatGPT three times at separate intervals, with sessions for each question at each attempt. To evaluate robustness, ChatGPT was challenged with an adversarial prompt after each answer choice on the third attempt, repeated three times within the same session. The model's confidence in its responses was also assessed by prompting it to rate its confidence from 1–10 on the third attempt and after each challenge prompt, all within the same session.

Overconfidence, Poor Repeatability and Robustness

Overall, while both GPT-3.5 and GPT-4 demonstrated reliable accuracy over time, their repeatability and robustness were poor. Additionally, both models exhibited overconfidence, with GPT-4 showing better insight into the likelihood of accuracy compared to GPT-3.5.

- Reliability: Both GPT-3.5 and GPT-4 displayed consistent accuracy across three attempts, with no significant difference observed between attempts for either model. GPT-3.5 achieved accuracies of 69.3%, 63.3%, and 60.7% across attempts, while GPT-4 achieved 80.6%, 78.0%, and 76.6%. Thus, both models demonstrated reliable accuracy over time.
- Repeatability: While the agreement in answer choices across attempts was poor for both models, GPT-4 exhibited higher repeatability compared to GPT-3.5. However, even for GPT-4, repeatability was suboptimal. Agreement between attempts varied from weak to strong for both models, with moderate agreement observed overall. Notably, consistent incorrect responses were more common for higher-order questions in GPT-3.5 but not in GPT-4.
- Robustness: Both GPT-3.5 and GPT-4 demonstrated poor robustness, frequently changing responses when challenged with an adversarial prompt. GPT-4 exhibited a higher propensity to alter responses compared to GPT-3.5. Additionally, successive challenge prompts resulted in a higher likelihood of response changes for both models, with GPT-4 consistently changing its response for all questions.
- Confidence: Both models frequently rated their confidence as high, even for incorrect responses, indicating overconfidence. However,

GPT-4 showed improved insight into the likelihood of accuracy, as it rated confidence lower for incorrect responses more frequently compared to GPT-3.5. Furthermore, after subsequent challenge prompts, GPT-4 consistently rated confidence lower compared to GPT-3.5.

Although GPT-4 showed improved repeatability compared to GPT-3.5, both models were susceptible to changing responses when challenged. This lack of adversarial resistance may stem from prioritising natural language production over accuracy. Despite GPT-4 showing slightly better insight into response accuracy, both models were frequently overconfident, indicating the need for caution in relying on their confidence ratings. While these default versions of ChatGPT have potential for clinical and patient-facing applications, their limitations suggest that optimization, including parameter adjustment and guardrails, is necessary for radiology-specific tasks to ensure reliability, repeatability, and robustness.

Source: RSNA Radiology

Image Credit: iStock

Published on : Thu, 23 May 2024