HealthManagement
Promoting Management and Leadership

Accuracy and Reliability of AI-Generated Educational Pamphlets for Radiology

Large language models (LLMs) are advanced deep learning models that generate human-like text using a multilayer neural network architecture. Trained unsupervised on vast amounts of text data, they understand word relationships and predict subsequent words in a sequence. ChatGPT, developed by OpenAI and launched on November 30, 2022, is an LLM based on GPT-3.5 and GPT-4 models. It processes entire input texts and integrates human feedback to better align outputs with user intent, enabling it to generate conversational text, translate languages, create diverse content, and answer a wide range of questions.

Despite their capabilities, LLMs like ChatGPT face challenges, particularly in healthcare, due to the risk of generating inaccurate information or "hallucinations"—responses not grounded in factual data. Inaccuracies in healthcare can disrupt clinical workflows and endanger patient health, making it essential to verify LLM outputs for accuracy and reliability.

LLMs have the potential to enhance healthcare communication between providers and patients. Studies have assessed the accuracy and consistency of LLM responses to healthcare-related questions, finding that ChatGPT can correctly answer up to 70.8% of questions related to lung cancer screening and prevention in radiology reports. However, LLMs can struggle with complex or technical medical terminology.

Patient education, especially in Interventional Radiology (IR), is often hindered by patients' lack of understanding. Accurate and clear information is crucial for informed decision-making and optimal outcomes. LLMs can help bridge the communication gap between medical professionals and patients by providing accurate and safe information.

This study published in Academic Radiology evaluates the accuracy and reliability of educational pamphlets created by ChatGPT for common IR procedures, aiming to explore their clinical relevance and practical utility in the field.

100 Pamphlets for 20 IR Procedures

Three radiologists selected twenty frequently performed interventional radiology (IR) procedures to cover a broad range of techniques. The selection process involved using Google search terms like "most common interventional radiology procedures" and consulting information from the Society of Interventional Radiology (SIR) and the Cardiovascular and Interventional Radiological Society of Europe (CIRSE). Five users (three radiologists and two radiologist trainees) independently instructed ChatGPT version 3.5 to generate patient education pamphlets for each selected procedure using identical prompts. This process resulted in 100 pamphlets, five for each of the 20 IR procedures. Two independent board-certified radiologists reviewed the generated pamphlets for structure, quality, and accuracy. They evaluated whether any important information was missing and identified potential errors, such as inaccurate or outdated content and hallucinatory information.

Pamphlets Structure and Evaluation

For each of the 20 different IR procedures, five sets of pamphlets were generated, resulting in a total of 100 educational pamphlets. Almost all pamphlets followed a consistent structure, including:

- Introduction: Brief description of the procedure and clinical rationale.
- Pre-procedural Preparation: Steps patients should take before the procedure.
- Procedure Description: What to expect during and after the procedure.
- Outcomes and Follow-up: Expected results and follow-up care.
- Risks an-d Complications: Potential risks and complications.

- Conclusion: Encouragement to contact healthcare providers with further questions.

In evaluating the pamphlets, shortcomings were identified in at least one pamphlet within 11 of the 20 groups. The other 9 groups had no major inaccuracies or shortcomings. Overall, 30% (30/100) of the pamphlets had shortcomings, totaling 34 specific inaccuracies:

- Missing information about sedation (10 instances)
- Inaccuracies related to procedural complications (8 instances)
- Lack of details on post-procedural precautions (6 instances)
- Insufficient information on pre-procedural preparation (5 instances)
- Incomplete outline of procedural steps for liver cancer Y-90 radioembolization (5 instances)

No hallucinatory content or fabricated data were found in any pamphlets. Key-word co-occurrence networks were used to compare the content of the pamphlets. Consistency was observed across each group of pamphlets, with similar themes and connections of words.

Accuracy, Challenges, and Future Directions

While ChatGPT generated structured pamphlets, 30% of them contained inaccuracies, such as missing sedation information, procedural complications, and pre-procedural preparation details. These omissions could pose safety risks, including uninformed consent and procedural delays. Accurate information on procedural steps and precautions, especially for complex procedures like Y-90 Trans-Arterial Radioembolization, is crucial for patient safety. Strategies are needed to mitigate these inaccuracies.

ChatGPT has received mixed opinions regarding its benefits versus risks. Although it can enhance output, there are concerns about hallucinations causing misinformation. This study did not observe hallucinations, but previous studies have reported inaccuracies in ChatGPT's responses to healthcare questions. For instance, ChatGPT responses to lung cancer screening questions were incorrect 17.5% of the time. Despite this, users preferred ChatGPT responses over physicians' in 78.6% of cases for public social media healthcare questions.

ChatGPT faces several challenges, including the need for up-to-date medical data. It is not fine-tuned on current clinical guidelines or radiologic terminology, potentially leading to outdated or inaccurate advice. This study is pioneering in evaluating ChatGPT’s efficacy in creating educational materials for IR procedures. The evaluation found consistent keyword themes but identified variability in structure and style across pamphlets, which could impact patient comprehension.

While ChatGPT has potential for creating educational medical content, it currently lacks complete accuracy and introduces structural variabilities. Ongoing human supervision and expert validation are essential. Future advancements in large language models may enhance reliability and utility in medical content creation. This study highlights the potential for expanding ChatGPT’s role in patient education and calls for further exploration of its ability to tailor content to individual patient needs.

Source: Academic Radiology

Image Credit: iStock

Published on : Tue, 11 Jun 2024