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

Preliminary Evaluation of the State-of-the-art Large Language Models in Processing Reports from the American Association of Physicists in Medicine

Hossein Jafarzadeh, MSc, PhD Candidate, McGill University; Jonathan Kalinowski, MSc; Farhood Farahnak, PhD, MSc; Shirin A. Enger, PhD, MSc

Introduction/Background

Reports from the American Association of Physicists in Medicine (AAPM) contain consensus guidelines and tabulated reference data essential for daily clinical tasks in radiology and radiotherapy. A chatbot capable of accurately answering questions regarding the reports would facilitate compliance with the AAPM guidelines. Retrieval-Augmented Generation (RAG) allows large language models (LLMs) to find the answer to questions from large amounts of text due to their context comprehension, attention mechanisms, and reasoning abilities. We evaluated Google's Gemini 1.5 Pro and OpenAI's GPT 4O on answering technical questions from two AAPM reports using human evaluation.

Methods/Intervention

Out of 259 AAPM reports, reports number 233 (TG233) and 084S were chosen for system evaluation, totaling 90 PDF pages. The PDFs were converted to text, and tables and images were extracted using the available APIs for each system. For each report, 5 technical questions were designed, and the models were asked to find the correct answers within the text. Finally, two graduate students with accredited training in medical radiation physics evaluated the models' responses and scored them from 1 to 5 based on accuracy and conciseness.

Results/Outcome

Gemini and ChatGPT scored 3.7 ± 1.4 and 2.7 ± 1.3 out of 5, respectively, in the human evaluation, showing Gemini's superiority. Gemini's responses averaged 144 ± 72 words, shorter and more concise than ChatGPT's 233 ± 87 words, though both were much longer than the ground truth (40 ± 14 words). Human evaluators noted that both models' answers were often verbose and inaccurate when questions required an understanding of relevant physics.

Conclusion

In conclusion, this experiment demonstrates LLMs' capability to understand AAPM reports and answer related questions. Future work will include integrating a search module to retrieve relevant reports for queries. Additionally, training task-specific LLMs, such as those fine-tuned on medical physics textbooks, is essential. A robust evaluation framework is also necessary to accurately assess these systems.

Statement of Impact

While assessing the capability of commercial LLMs in processing reference documents specific to the medical physics domain, this work signifies the need for a more standardized method of evaluating model performance on technical reference documents.

Keywords

Artificial intelligence; Large Language Models; Medical Physics; Computer Tomography; Quality Assurance; Chatbot