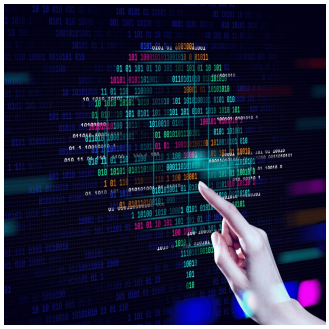
HealthManagement
Promoting Management and Leadership

Evaluating the Reliability of Large Language Models in Biomedical Knowledge



The use of Large Language Models (LLMs) in biomedical research has the potential to revolutionise the accessibility of scientific knowledge. With LLMs like ChatGPT and GPT-4 showing promising linguistic capabilities, there is growing interest in leveraging these models for tasks like chemical compound identification and understanding relationships between biological entities. However, a big challenge remains: ensuring that the information generated by these models is factual and accurate. A recent article published in the Journal of Biomedical Informatics proposes criteria to assess LLMs' biomedical knowledge capabilities, such as fluency, prompt alignment, semantic coherence, and factual knowledge.

Framework for Evaluation

The framework proposed for assessing LLMs in biomedical contexts is centred on three sequential steps: fluency, factuality, and specificity. Initially, non-experts evaluate the model's outputs for linguistic fluency, ensuring the generated text is syntactically correct and semantically coherent. If these criteria are met, the output is then reviewed by experts for factual alignment with domain-specific knowledge. The third step involves assessing the specificity of responses to ensure they align with the level of abstraction required by the prompt. This tiered approach enables a more efficient evaluation process by filtering out irrelevant outputs early on, reducing the workload on domain experts.

Performance of LLMs on Biomedical Tasks

To understand how well LLMs encode biomedical knowledge, eleven state-of-the-art models were evaluated on two tasks: generating chemical compound definitions and determining chemical-fungus relationships. Despite advancements in model fluency and linguistic capabilities, the ability of LLMs to generate accurate biomedical knowledge remains limited. The models often generated factually incorrect information, with some hallucinating information entirely. Additionally, biases towards overrepresented entities like Aspergillus were prevalent, indicating that these models rely on the most frequently encountered information in their training data. Even specialised models, such as BioGPT, struggled with factuality, particularly when context was added to prompts.

Variability Across Models and Impact of Prompt Design

The main finding from this evaluation is the variability in performance across different models and prompt designs. While models like GPT-4 and ChatGPT demonstrated the highest fluency and coherence, they frequently failed in factual accuracy and specificity, often generating plausible but incorrect information that would be difficult for non-experts to detect. Conversely, models specifically trained on biomedical corpora, such as BioGPT and its larger variant, showed improved domain specificity but were still prone to factual errors and biases. Interestingly, model size impacted performance, with larger models like Llama 2-70B performing better in generating correct chemical-fungus relations than their smaller counterparts.

While LLMs exhibit significant potential in assisting biomedical research, their current limitations in encoding and generating factual knowledge are evident. The framework for expert evaluation proposed in this article highlights the importance of careful assessment of LLM-generated content, particularly in high-stakes fields like biomedicine where accuracy is paramount. The models' biases towards common entities and their tendency to hallucinate underscore the need for more robust training and evaluation mechanisms. Future advancements should focus on enhancing domain specificity, improving factual accuracy, and developing systematic approaches for prompt optimisation to maximise the potential utility of LLMs in biomedical research.

Source: Journal of Biomedical Informatics
Image Credit: iStock

Published on : Mon, 7 Oct 2024

© For personal and private use only. Reproduction must be permitted by the copyright holder. Email to copyright@mindbyte.eu.