Language model "UroBot“ surpasses the accuracy of experienced urologists

Scientists at the German Cancer Research Center (DKFZ), together with doctors from the Urological Clinic of the Mannheim University Hospital, have developed and successfully tested a chatbot based on artificial intelligence. "UroBot" was able to answer questions from the urology specialist examination with a high degree of accuracy, surpassing both other language models and the accuracy of experienced urologists. The model justifies its answers in detail based on the guidelines.

With advances in personalized oncology, urological guidelines are becoming increasingly complex. Whether in the tumor board, on the ward or in the practice, a precise second-opinion system for medical decisions in urology could support doctors in evidence-based and personalized care, especially when time or capacity is limited.

Large language models (LLMs) such as GPT-4 have the potential to retrieve medical knowledge and answer complex medical questions without additional training. However, their applicability in clinical practice is often limited due to outdated training data and a lack of explainability. To overcome these hurdles, a team led by Titus Brinker of the DKFZ developed "UroBot", a specialized chatbot for urology that was supplemented by the current guidelines of the European Society of Urology.

UroBot is based on OpenAI's most powerful language model, GPT-4o. It uses a customized method of retrieval-augmented generation (RAG) that is able to retrieve relevant information from hundreds of documents in a targeted manner in response to the individual question in order to provide precise and explainable answers. The modified model was tested on 200 specialist questions from the European Board of Urology and evaluated in several rounds.

UroBot-4o answered questions on the specialist examination correctly 88.4 percent of the cases, outperforming the most up-to-date model GPT-4o by 10.8 percentage points. This means that UroBot not only outperforms other language models, but also exceeds the average performance of urologists in the specialist examination, which is reported in the literature as 68.7 percent. In addition, UroBot shows a very high degree of reliability and consistency in its answers.

UroBot's answers can be verified by clinical experts, since the software identifies the decisive sources and text sections: "The study shows the potential of combining large language models with evidence-based guidelines to improve performance in specialized medical fields. The verifiability and the very high accuracy at the same time make UroBot a promising assistance system for patient care."The use of comprehensible language models like UroBot will become extremely important in patient care in the next few years and will help to ensure guideline-based care across the board, even as therapy decisions become increasingly complex," says Brinker.

The research team has published the code and instructions for using UroBot to enable future developments in urology, as well as in other medical fields.

Publication:

Martin J. Hetz, Nicolas Carl, Sarah Haggenmüller, Christoph Wies, Jakob Nikolas Kather, Maurice Stephan Michel, Frederik Wessels, Titus J. Brinker: **Superhuman Performance on Urology Board Questions Using an Explainable Language Model Enhanced with European Association of Urology Guidelines.**

Press release

09-Oct-2024
Source: German Cancer Research Center (DKFZ)

Further information

- German Cancer Research Center (DKFZ)