UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Seminar Announcement

DIPARTIMENTO DI INFORMATICA

The Seminars on "Information Technology Outlook" - PhD Program in Computer Science and Mathematics

Shlomo Berkovsky
Professor
Interactive Medical AI research stream
Australian Institute for Health Innovation
Macquarie University, Sydney, Australia

Thursday June 26, 2025
15:00

Aula Gödel
Dept. of Computer Science

Expert Evaluation of LLMs for Clinical Dialogue Summarisation

In this work we assessed the performance of LLMs summarising clinical dialogues using computational metrics and human evaluations. We conducted an exploratory evaluation of five models: one general summarisation model, one fine-tuned for general dialogues, two fine-tuned with anonymised clinical dialogues, and ChatGPT. These models were assessed using the ROUGE and UniEval metrics, and expert human evaluation was done by clinicians comparing the generated summaries against a clinician generated summary (gold standard). The fine-tuned transformer model scored the highest when evaluated with ROUGE, while ChatGPT scored the lowest overall. However, using UniEval, ChatGPT scored the highest across all the evaluated domains. Similar results were obtained when the systems were evaluated by clinicians. Statistical analyses showed superiority of ChatGPT's and human summaries over the summaries generated by other models. These results indicated that the performance of ChatGPT in summarising clinical dialogues approached the quality of human summaries. The study also found that the ROUGE metrics might not be reliable for evaluating clinical summaries, whereas UniEval correlated better with human ratings. Further evaluation using diverse clinical dialogues is needed to verify the reliability and safety of LLM generated summaries.

Shlomo Berkovsky *is the leader of the Interactive Medical AI research stream at Macquarie University. The stream focusses on the use of Artificial Intelligence and Machine Learning methods to develop usable patient models and personalised predictions of diagnosis and care. The stream also studies how clinicians and patients interact with health technologies and how Large Language Models can improve patient care. His other areas of expertise include user modelling, online personalisation, and behaviour change technologies.*