Performance of ChatGPT-4o vs ChatGPT-5.2 in summarizing adverse drug reactions

Rogier W.M.A. van der Zanden[1], Marielle Slikkerveer[1], Receb Gündogan[2], Alan Abdulla[2], Fatma Karapinar-Çarkit[1,3] Johanna H.M. Driessen[1,3]

1 Department of Clinical Pharmacy & Toxicology, Maastricht University Medical Center+, Maastricht, The Netherlands
2 Department of Hospital Pharmacy, Erasmus University Medical Center, Rotterdam, the Netherlands
3 CARIM, Department of Clinical Pharmacy, Cardiovascular Research Institute Maastricht, Maastricht University, Maastricht, The Netherlands

Background and importance

Hospital pharmacists frequently assess adverse drug reactions (ADRs) in patients, a task that is time consuming and requires expert knowledge, especially in patients with polypharmacy. Large language models (LLMs) such as ChatGPT may accelerate the assessment of ADRs by generating a summarizing overview. With each new release of LLMs, performance is expected to improve, but systematic evaluation is lacking.

Aim and objectives

The aim of this study was to investigate the performance of the new ChatGPT-5.2 model compared to ChatGPT-4o in producing an accurate overview of ADRs.

Materials and methods

Thirty commonly prescribed cardiovascular medications were analyzed using an engineered single prompt restricting both models to a national drug database[1]. Across five iterations per medication (150 runs in total), each model extracted ADRs and categorized them as common (1–10%) or very common (>10%). Outputs were manually validated by two persons. Accuracy (correct ADRs vs database), hallucinations (false ADRs), and omission errors (missed ADRs) were calculated. Descriptive statistics were used, mean accuracy and mean errors between ChatGPT versions were compared using independent t-tests. Proportion correct runs (no mistakes in one iteration for one medication), and correct across all runs (no mistakes in five iterations for one medication) were compared using an independent z-test.

Results

ChatGPT-5.2 substantially outperformed ChatGPT-4o. Accuracy of correctly summarizing ADRs and frequency was better when using ChatGPT-5.2 compared to ChatGPT-4o. Most common errors were omissions. Hallucination errors did not occur in ChatGPT-5. Results are shown in table 1.

Table 1 Performance of ChatGPT-4o vs. ChatGPT-5.2

	ChatGPT-4o % (SD)	ChatGPT-5.2% (SD)
Accuracy	85.9 (28.1)	97.6 (12.3)
Omissions	23.3 (31.9)	2.5 (12.4)
Hallucinations	9.5 (20.7)	0 (0)
Correct runs	61/150 (40.7)	140/150 (93.3)
Medications correct across all runs	4/30 (13.3)	22/30 (73.3)

SD: standard deviation. p<0.001 for all comparisons

Conclusion and relevance

ChatGPT-5.2 clearly outperformed ChatGPT-4o in summarizing ADRs of the tested medications. Omission errors were the most frequently occurring errors. Hallucinations were not observed with ChatGPT-5.2. These findings suggest that ChatGPT-5.2 should be further evaluated as a tool for hospital pharmacists in detecting ADRs. Future research should address other medication groups and evaluate real-world clinical application.

References

[1] Zorginstituut Nederland. Farmacotherapeutisch Kompas. Available through https://farmacotherapeutischkompas.nl.

R. Van der Zanden, r.vander.zanden@mumc.nl

Maastricht UMC+ Erasmus MC Maastricht University

5PSQ-078