

Evaluating the Reliability of ChatGPT for Health-Related Questions: A Systematic Review

Mohammad Beheshti, MS, Imad Eddine Toubal, PhD, Khuder Alaboud, PhD, Mohammed Almalaysha, MS, Olabode B. Ogundele, MS, MA, Hamza Turabieh, PhD, Nader Abdalnabi, MBA, MS, Suzanne A. Boren, PhD, Grant J. Scott, PhD, Butros M. Dahu, PhD

Supplementary Materials

Table 1: The 6 out of the 128 studies which reported specific measures such as precision, recall, and specificity.

Authors	Year	Precision	Recall	Specificity
Sarbay, Ibrahim <i>et al.</i>	2023	0.39	0.57	0.34
Al-Ashwal, Fahmi Y <i>et al.</i>	2023	0.41	0.75	0.52
Gebrael, Georges <i>et al.</i>	2023	-	0.96	0.18
Coskun, Burhan <i>et al.</i>	2023	0.35	0.55	-
Benary, Manuela <i>et al.</i>	2023	0.23	0.31	-
Wilhelm, Theresa Isabelle <i>et al.</i>	2023	0.90	0.87	0.90

Table 1 shows six out of the 128 studies that reported specific measures such as Precision, Recall, and Specificity.

These measures, although informative, were inconsistently reported across studies, limiting the feasibility of presenting them.