Abstract Number: 6ER-037

PERFORMANCE AND CONCORDANCE OF ARTIFICIAL INTELLIGENCE IN THE BPS

D. GARCIA MARTINEZ[1], A.B. BENITO POUSADA[1], F. FERNANDEZ FRAGA[2], A. GONZALEZ FUENTES[1], M.R. MENGUAL BARROSO[1], I. GONZALEZ GARCIA[1], N. GARRIDO PEÑO[1], M. CARRERA SANCHEZ[1], L. FERNÁNDEZ VALENCIA[1], Y. MATEOS MATEOS[1], M. SEGURA BEDMAR[1].

[1]HOSPITAL UNIVERSITARIO DE MÓSTOLES, FARMACIA HOSPITALARIA, MADRID, SPAIN. [2]HOSPITAL 12 DE OCTUBRE, FARMACIA HOSPITALARIA, MADRID, SPAIN.

BACKGROUND AND IMPORTANCE

Artificial Intelligence (AI) is increasingly assuming a pivotal role in modern society. Its diverse applications are transforming numerous tasks, including those within hospital pharmacy. However, the development of robust AI evaluation tools is essential to ensure their effective integration into professional workflows.

AIM AND OBJECTIVES

To assess the performance and concordance of three AI systems (ChatGPT 3.5, ChatGPT 4.0, and Gemini) in addressing Board of Pharmacy Specialties (BPS) examination questions.

MATERIAL AND METHODS

Observational and cross-sectional study conducted in August 2024. All sample questions and answers provided on the BPS website, designed to familiarize candidates with the structure and format of BPS certification exams, were extracted. A protocol was developed to guide the AIs in responding to the questions, instructing them to rely on high-quality references and to refrain from generating answers not based on data.

Tests Conducted: 3 tests per AI. Each test administered by a different researcher. In cases of uncertainty, AIs encouraged to select "DK/NR" (Doesn't Know/No Response).

Chi-Square test: Comparing proportions of correct answers.

Kappa index & Altman's criteria: Assessing concordance between AI responses and researchers.

RESULTS

137 questions

AI Model	Test 1 (%)	Test 2 (%)	Test 3 (%)	Mean (%)
ChatGPT 3.5	83.2	76.6	83.9	**81.3**
ChatGPT 4.0	86.1	83.9	73.7	**81.3**
Gemini	65.0	59.1	65.0	**63.0**

AI Model	Test 1 (Kappa)	Test 2 (Kappa)	Test 3 (Kappa)	Mean (Kappa)	Agreement Level
ChatGPT 3.5	0.773	0.862	0.792	**0.809**	Excellent Agreement
ChatGPT 4.0	0.686	0.941	0.676	**0.809**	Excellent Agreement
Gemini	0.548	0.621	0.584	**0.572**	Moderate Agreement

Statistically significant differences were found by ChatGPT 4.0 and ChatGPT 3.5 (81.3%) compared to Gemini (63.0%) ($p < 0.01$).

No statistically differences were found between ChatGPT 3.5 and 4.0.

CONCLUSION AND RELEVANCE

ChatGPT 3.5 and 4.0 show comparable performance with excellent agreement, while Gemini has significantly lower accuracy and consistency.