

## WOULD CHATGPT PASS THE RESIDENT INTERNAL PHARMACIST EXAM?

C. GONZÁLEZ<sup>1,2</sup>, J. NIETO DE VICENTE<sup>2</sup>, M. TORREGO ELLACURÍA<sup>2</sup>, G. HERNANDO LLORENTE<sup>1</sup>, P. PASTOR VARA<sup>1</sup>, M.FERNÁNDEZ-VÁZQUEZ CRESPO<sup>1</sup>, J. CORAZÓN VILLANUEVA<sup>1</sup>, L. LLORENTE SANZ<sup>2</sup>, M. LUACES<sup>2</sup>, M.T. BENÍTEZ GIMÉNEZ<sup>1</sup>. Contact: cristina.gonzalez.perez@salud.madrid.org <sup>1</sup>HOSPITAL PHARMACY, HOSPITAL CLINICO SAN CARLOS; IDISSC, MADRID, SPAIN. <sup>2</sup>INNOVATION UNIT; IDISSC, MADRID, SPAIN.

#### **Background and importance**

Assessing ChatGPT's performance in the Health Training exam for Pharmacy specialization (FIR) holds significance ingauging Al's role in healthcare education.

### Aim and objectives

To assess ChatGPT's ability to respond to and potentially pass the Health Training exam for Pharmacy specialization (FIR).



#### Materials and methods

A multidisciplinary team consisting of hospital pharmacists, physicians and biomedical engineers selected an exam version for the 2022 session.

One question was excluded due to the presence of an image.

A brief introduction, providing context about the FIR exam and its contents, was added at the beginning of the conversation. Correct answers received a value of 3 points, while incorrect ones subtract incurred a deduction of 1 point. In the 2022 call, a minimum score of 97 points was necessary to be eligible for allocation of FIR positions.

ChatGPT's performance, defined as the percentage of correct answers, was evaluated through three different approaches:



# Manual introduction into the OpenAl interface

Two sets of 50 randomly selected questions were manually input into the OpenAl web interface during the same conversation.

# Through Application Programming Interface (API)

A total of 209 questions, including both questions and their four possible answers were solved by API for Python from a spreadsheet.

# Using Natural Language Processing (NLP)

Open-ended questions lacking predefined possible answers were extracted by API for Python, followed by the application of NLP.

NLP assessed the similarity between APIgenerated responses and actual responses, providing a more accurate evaluation of ChatGPT's human-like performance in a multiple-choice exam. The similarity metric compared feature vectors of sentences and generated a value representing the degree of similarity, with a maximum value of 1 signifying

a perfect match and thus a correct answer.

# **60 & 66% 45.5-49% 50.2-52.6%**

#### Results

Using the manual inclusion method, we achieved 60% and 66% accuracy in 50 randomly selected questions (score equivalent to 280 and 328 points, respectively). The second method yielded a success a success rate of 45.5 to 49.0%, equating to 164-192 points. In the third method, values of 50.2-52.6% (200-220 points) were obtained.

#### **Conclusion and relevance**

The findings demonstrate ChatGPT's variable ability to provide correct responses to FIR questions depending on the methodology employed. Regardless of the approach, ChatGPT consistently achieved the minimum score required for participation in the allocation of FIR positions.