ЧИ ЗДАТНІ ЧАТ-БОТИ ІЗ ГЕНЕРАТИВНИМ ШТУЧНИМ ІНТЕЛЕКТОМ УСПІШНО СКЛАСТИ ЕКЗАМЕНАЦІЙНЕ ТЕСТУВАННЯ ДЛЯ АТЕСТАЦІЇ ЛІКАРІВ-ПУЛЬМОНОЛОГІВ?

В. Л. Побережець, І. О. Радогощин

Вінницький національний медичний університет ім. М. І. Пирогова, Вінниця, Україна

Резюме. Чат-боти із генеративним штучним інтелектом (ШІ) за досить короткий проміжок часу (із 2022 року) інтегрувались у всі сфери нашого життя, навіть якщо ми цього не помічаємо. Медицина та її окремі галузі, такі як пульмонологія, не стала виключенням і генеративний ШІ відмінно проявив свій потенціал у інтерпретації візуалізаційних методів досліджень, поясненні результатів спірометрії, допомозі у прийнятті клінічних рішень та навчанні. Однак досі залишається дискутабельним питання чи здатні моделі із генеративним ШІ наблизитись до результатів живих лікарів у офіційному медичному ліцензійному тестуванні.

Мета роботи: Оцінити здатність чат-ботів із генеративним штучним інтелектом у вирішенні екзаменаційного тестування для атестації лікарів-пульмонологів.

Матеріали та методи: У грудні 2024 року нами було запропоновано вирішити екзаменаційні тести із бази запитань для атестації лікарів-пульмонологів трьом найпоширенішим в Україні безкоштовним чат-ботам із генеративним ШІ — ChatGPT версія 3.5, Microsoft Copilot та Gemini. Даним чат-ботам було представлено завдання вирішити 1095 тестових завдань із загальної бази даних, після чого було здійснено аналіз відповідей на запитання про бронхіальну астму (92 запитання) та алергопатологію (35 запитань).

Результати: Точність ChatGPT у вирішенні пульмонологічних тестів склала 95 % (n = 1037 правильних відповідей), Microsoft Copilot — 92 % правильних відповідей (n = 1008), а Gemini — 81 % правильних відповідей (n = 890). У відповідях на запитання, що стосувались діагностики та лікування алергопатології найкращі результати показав Microsoft Copilot із 100 % правильних відповідей (n = 35); ChatGPT набрав 94,3 % правильних відповідей (n = 33), Gemini — 85,7 % правильних відповідей (n = 30). На запитання про бронхіальну астму ChatGPT відповів правильно у 91,3 % випадків (n = 84), Gemini — 79,4 % (n = 73), Copilot — 89,1 % (n = 82). Усі чат-боти показали кращі результати при відповіді на запитання, що мали єдину правильну відповідь ніж на запитання із множинними правильними відповідями: ChatGPT — 92,9 % проти 75 %, Gemini — 83,3 % проти 37,5 %, Copilot — 94 % проти 37,5 % правильних відповідей.

Висновки. Наше дослідження встановило, що чат-боти із генеративним ШІ продемонстрували високу результативність у вирішенні екзаменаційного тестування для атестації лікарів-пульмонологів, що можна вважати прохідним для лікаря-спеціаліста. Зокрема, це стосується і запитань щодо бронхіальної астми та алергопатології. Найкращий загальний результати продемонстрував ChatGPT, який правильно відповів на 95 % усіх тестів. Було виявлено, що генеративний ШІ значно краще справлявся із вирішенням запитань із єдиною правильною відповіддю порівняно із запитаннями із множинними правильними відповідями.

Ключові слова: велика мовна модель, штучний інтелект, ChatGPT, бронхіальна астма, алергопатологія.