УДК 004.05+004.8 https://doi.org/10.32515/2664-262X.2025.12(43).2.27-35

О. В. Присяжнюк, доц., канд. техн. наук, **А. В. Пузікова**, доц., канд. фіз.-мат. наук, **Д. А. Оришечко**
Центральноукраїнський державний університет імені Володимира Винниченка, м. Кропивницький, Україна
e-mail: elena_drobot@ukr.net, a.v.puzikova@cuspu.edu.ua, 12026083@cuspu.edu.ua

Експериментальне дослідження ефективності моделі GPT-4o для оцінювання якості користувацьких інтерфейсів із врахуванням безпекових ризиків

У дослідженні вивчається ефективність мультимодальної моделі GPT-4o для використання її в ролі цифрового експерта з якості користувацького інтерфейсу з урахуванням безпекових ризиків шляхом проведення порівняльного аналізу оцінок, отриманих від GPT-4o та реальних експертів в єдиній системі критеріїв та UX-показників. Експеримент на вибірці з 20 сайтів ЗВО показав високий рівень відповідності між оцінками, отриманими від моделі GPT-4o та висновками фахівців, що підтверджено статистично значущими коефіцієнтами узгодженості оцінок. Виявлено, що модель GPT-4o здатна виявляти UX-вразливості, зокрема у сфері доступності, які можуть залишатися поза увагою експертів, що підтверджується аналітичними звітами. Отримані результати підтверджують доцільність використання мультимодальних моделей ШІ як засобів підсилення експертного аналізу та зменшення ресурсних витрат у процесах аудиту інтерфейсів.

аналіз вимог інтерфейсів, якість користувацьких інтерфейсів, GPT-4o, експертні оцінки, безпекові ризики

Постановка проблеми. Питання контролю якості користувацького інтерфейсу (User interface, UI) охоплюють широкий спектр заходів, що у сукупності забезпечують ефективну роботу вебпродукту, комфортну і безпечну взаємодію користувача із додатком та підвищення його конкурентоспроможності. Водночас аудит UI – процес тривалий і ресурсозатратний, і для малих команд він створює надмірне функціональне навантаження, потребуючи додаткових часу, зусиль та ресурсів. Зокрема, найбільш затратними по часу і по ресурсам є оцінювання якості інтерфейсу з точки зору користувацького досвіду (User Experience, UX) потенційних користувачів. Тому спостерігається зростання уваги до інтеграції інструментів штучного інтелекту (ШІ) у процеси контролю якості інтерфейсів з метою їх автоматизації, як у наукових колах дослідників, так і професійних колах практиків. [1, 2].

Сучасні інструменти автоматизації аудиту якості інтерфейсів з використанням ШІ в основному направлені на перевірку технічної продуктивності та тестування функціональності вебпродуктів [3, 4], але вони не можуть охопити когнітивні та емоційні аспекти користувацького досвіду, особливо при оцінюванні того, як користувачі сприймають та взаємодіють з інтерфейсами на психологічному рівні [5] і які вразливості інтерфейсу можуть обумовити потенційні безпекові ризики. Поява та бурхливий розвиток великих мовних моделей (LMM) з просунутими можливостями розуміння зображень, таких як мультимодальна модель GPT-4o, являють собою потенційну можливість організації автоматизованого оцінювання якості користувацького інтерфейсу з погляду зручності, доступності та інших важливих параметрів UX, які забезпечують комфортну та безпечну взаємодію користувача із продуктом. Але при цьому важливе розуміння, наскільки ефективно модель GPT-4o може функціонувати в якості експерта-оцінювача, уловлюючи аспекти взаємодії та безпекові ризики інтерфейсу з позицій користувацького досвіду. Зазначені вище обставини

зумовили інтерес до дослідження в цій галузі.

Аналіз останніх досліджень і публікацій. В останні роки дослідниками активно ведеться наукова розвідка інтеграції інструментів LMM в процеси оцінювання якості користувацького інтерфейсу [5, 6] та безпекових ризиків [7-9].

Дослідження щодо застосування інструментів ШІ для аналізу якості інтерфейсів підтверджують ефективність таких підходів для проведення ретельного оцінювання UX/UI. Інструменти ШІ здатні генерувати змістовні аналітичні висновки щодо якості інтерфейсу, послідовно виявляти проблемні елементи дизайну та давати рекомендації для їх усунення на рівні з традиційними методами оцінювання [5, 6, 10]. Поява мультимодальних моделей з потужними навичками візуального сприйняття та розуміння, таких як GPT-4o відкриває нові можливості для автоматизації процесів оцінювання якості інтерфейсів вебпродуктів. Дослідження [11] присвячено питанням застосування моделі GPT-4o для оцінювання зручності інтерфейсів вебпродуктів на основі евристик Нільсена. Авторами проводилось дослідження продуктивності GPT-4o шляхом порівняння результатів евристичної оцінки інтерфейсів, отриманої від експертів, з результатами отриманих за допомогою GPT-4o, яке виявило, що GPT-4o виявляє проблеми зручності та присвоює їм оцінки серйозності, аналогічні експертним, але при цьому йому не вистачає глибини та гнучкості. У публікації [12] аналізувались можливості використання LMM для отримання синтетичних евристичних оцінок якості користувацького інтерфейсу мобільних додатків. Порівняльний аналіз синтетичних евристичних оцінок з еталонним набором оцінок, отриманих від людей показав здатність LMM оцінювати компоненти користувацького інтерфейсу і підтвердив її потенціал для генерації якісних даних, аналогічних оберненому зв'язку від користувача.

Оскільки процес розробки UX-дизайну тісно пов'язаний із впровадженням заходів безпеки з урахуванням потреб, можливостей та поведінки користувачів, в останні роки питання балансу між безпекою та зручністю використання ІТ-продукту обговорюється як в науково-дослідних роботах, так і серед професіоналів-практиків. Так, у статті [7] розглядається аспект доступності (accessibility) в безпеці, а саме: як спрощені/альтернативні інтерфейси можуть ослабити заходи безпеки; описуються складні виклики, з якими стикаються розробники під час покращення доступності та інклюзивності рішень у сфері кібербезпеки; надаються рекомендації з дизайну інклюзивних механізмів тощо. Дослідженню візуальних аспектів кібербезпеки, тобто впливу естетичної інформації на сприйняття й довіру користувачів до програмного продукту, присвячена робота [8]. UX-дизайнери також мають враховувати ризики фішингових атак під час проєктування користувацьких інтерфейсів. У статті [9] наведено практичні рекомендації для розробників поштових клієнтів щодо запобігання таким атакам через відповідні рішення в UX-дизайні.

Огляд публікацій показує, що питання розробки та дослідження нових інструментів для інтегрування ШІ у процеси оцінювання якості інтерфейсів є вельми актуальними та перспективними. Це дозволить значно знизити фінансові та часові витрати на тестування та підвищити якість програмних продуктів. Разом з цим, досліджуючи ефективність імплементації ШІ в автоматизовану систему оцінювання якості користувацького інтерфейсу, слід враховувати можливі безпекові ризики UX-дизайнів. Аналіз наявних у відкритому доступі джерел свідчить про недостатність приділеної уваги напрямку інтеграції інструментів LLM у процеси оцінювання якості користувацького інтерфейсу із врахуванням безпекових ризиків.

Постановка завдання. Мета роботи – дослідити ефективність мультимодальної моделі GPT-4o як інструмента автоматизованого оцінювання якості користувацьких інтерфейсів із урахуванням безпекових ризиків.

Для досягнення поставленої мети слід вирішити такі задачі: 1) проаналізувати систему критеріїв оцінювання якості користувацьких інтерфейсів та визначити притаманні їй безпекові ризики, що можуть впливати на результати експертного аналізу; 2) розробити процедуру порівняння експертних оцінок і результатів GPT-4o в єдиному просторі критеріїв та шкал, придатному для подальшого статистичного опрацювання; 3) провести емпіричне дослідження на вибірці 20 сайтів ЗВО, зібравши оцінки експертів-людей та результати моделі GPT-4o відповідно до розробленої процедури; 4) виконати статистичний аналіз узгодженості оцінок, зокрема обчислення коефіцієнтів узгодженості між оцінками експертів та GPT-4o; 5) порівняти аналітичні звіти про виявлені безпекові ризики, отримані від експертів та GPT-4o.

Виклад основного дослідження. Для перевірки якості інтерфейсу з позиції користувацького досвіду (UX-оцінювання) зазвичай застосовуються евристики Нільсена: видимість статусу системи, зв'язок між системою та реальним світом, контроль і свобода користувача, узгодженість і стандарти, запобігання помилкам, розпізнавання замість запам'ятовування, гнучкість та ефективність, естетика та мінімалізм візуального дизайну, допомога у виявленні та виправленні помилок, а також довідка та документація. У деяких дослідженнях вказується на використання інтегрованих евристик Нільсена та/або базових критеріїв взаємодії користувача з інтерфейсом [5, 11]. В нашому дослідженні експертне оцінювання якості UX інтерфейсу відбувалось за чотирма критеріями (Таблиця 1). Наведені критерії охоплюють базові аспекти взаємодії з користувачем, такі як інтуїтивно зрозуміле та зручне сприйняття екрану, чіткість та доступність інтерфейсу, візуальний дизайн та структурованість контенту.

Таблиця 1 - Критерії оцінювання якості UX інтерфейсів

№	Критерій	Що оцінюється	Показники
1	зручність	Наскільки інтерфейс зручний для користувача та не викликає зайвих труднощів	Логічність навігації, зрозумілість елементів, відсутність помилок, швидкість освоєння
2	доступність	Чи може сторінку використовувати якомо ширше коло людей, включно з користу-вачами з обмеженими можливостями	Контрастність тексту, масштабованість, наявність альтернативного тексту для зображень
3	Візуальний дизайн	Естетика та візуальна привабливість інтерфейсу	Узгодженість кольорів і шрифтів, візуальна ієрархія, баланс елементів, приємність сприйняття
4	Інформаційна архітектура	Як упорядкована та структурована інформація на сторінці	Логічна структура меню, зрозуміле групування контенту, швидкий доступ до потрібних даних, мінімізація перевантаження інформацією

Джерело: розроблено авторами

З описаними вище критеріями оцінювання якості UX інтерфейсів пов’язані безпекові ризики для користувача.

Впровадження складних функцій безпеки збільшує ймовірність того, що користувачі припустять помилку або повністю відмовляться від захисту. Виходячи з принципу психологічної прийнятності, інтерфейси функцій безпеки повинні бути зручними і простими у використанні, щоб уникнути помилок користувача в їхніх застосунках. З іншого боку, помічено, що якщо емоційна реакція користувачів на візуальний дизайн є позитивною, це робить їх більш толерантними до незначних проблем із зручністю використання [9]. Таким чином забезпечення естетичності і зручності використання підвищує довіру користувача до вебпродукту і може сприяти зниженню безпекових ризиків з боку користувача. Критерій зручності використання

вебдодатку пов'язаний з таким ризиками, як втрата даних (причиною може стати непередбачуване або незручне розташування кнопок), помилкові дії користувача (наприклад, введення конфіденційних даних у неправильні поля і відправка форми), випадкове виконання критичних дій (причиною може стати схожий стиль інструментів для безпечних та ризикових дій) тощо.

Такий критерій оцінювання якості UX інтерфейсів як доступність слід розглядати як важливий вимір у сфері проєктування заходів безпеки, орієнтованих на людину [14]. Згідно з даними WebAIM (Web Accessibility in Mind), на лютий 2025 року близько 95% найпопулярніших вебсайтів світу недоступні для людей з обмеженими можливостями [15]. Рекомендації щодо підвищення доступності веб-контенту наведені у Керівних принципах доступності вебконтенту (WCAG) 2.1 [16] .

Помилкові дії користувача можуть бути наслідком як неправильно виконаного візуального дизайну так і нелогічно спроєктованої інформаційної архітектури. Недостатньо опрацьований візуальний дизайн інтерфейсу може підвищувати ризик фішингових атак. Цьому сприяють: відсутність в інтерфейсі впізнаваних візуальних елементів (логотипу, фірмової палітри, уніфікованого стилю), що спрощує створення зловмисниками підроблених сторінок, некоректна ієрархія кольорів робить критично важливі повідомлення малопомітними, недостатньо помітні попереджувальні елементи тощо. Таким чином, під час оцінювання якості користувацького інтерфейсу необхідно також перевіряти, чи враховано у дизайні безпекові ризики.

Автоматизація оцінювання якості користувацького інтерфейсу здійснювалась за допомогою мобільного додатку, розробленого на базі Flutter [17]. Цей застосунок реалізує відправку скріншотів інтерфейсів користувача для їх оцінки з використанням нейромережі, а також відображує отримані від моделі GPT-4o результати; проводить оцінювання інтерфейсу за заданими критеріями із врахуванням без пекових ризиків в бальній шкалі та формує текстовий аналітичний звіт про якість інтерфейсу і рекомендації щодо його покращення.

Для аналізу ефективності інтеграції GPT-4o в процеси оцінювання якості UX інтерфейсу з урахуванням безпекових ризиків, було проведено експеримент по порівнянню оцінок якості вебінтерфейсу 20 сайтів ЗВО, отриманих від GPT-4o, з аналогічними оцінками, отриманими від експертів. В якості експертів-оцінювачів було залучено п'ять фахівців-тестувальників з досвідом роботи від 3 до 6 років. Оцінки за кожним із запропонованих критеріїв (Таблиця 1) пропонувалось давати в загальноприйнятій п'ятибальній шкалі: 1 – низька якість, 2 – нижче середньої, 3 – середня якість, 4 – вище середнього, 5 – висока якість. Також експерти надавали короткі аналітичні звіти про якість інтерфейсу по кожному сайту, окремо по кожному критерію, та рекомендації щодо її покращення у випадку низької оцінки хоча би за одним з критеріїв із врахуванням безпекових ризиків.

Експеримент проходив у два етапи. На першому етапі порівнювались бальні оцінки якості інтерфейсів, отримані від моделі GPT-4o та в результаті проведення колективної експертизи.

Після збору відповідей експертів, числові дані були систематизовані і представлені матрицями вигляду:

$$A^k = \left\| a_{i,j}^k \right\|,\ k = \overline{1,4},\ i = \overline{1,20},\ j = \overline{1,5}, \qquad (1)$$

де $a_{i,j}^k$ оцінка і-го об'єкта (відповідного сайту ЗВО) j-м експертом за k-м критерієм.

Обробка даних колективної експертизи передбачала розрахунок усередненої оцінки об'єктів по кожному із критеріїв, в результаті чого критеріальні матриці (1) звелись до критеріальних векторів вигляду:

$$a^{(k)} = (\overline{a_i^k}),\ \ \overline{a_i^k} = \sum_{j=1}^{22} a_{ij}^k \Big/ 5, \tag{2}$$

де $\overline{a_i^k}$ середньозважена оцінка і-го об'єкта за k-м критерієм. Такі оцінки виражають, як правило, узгоджену думку групи, учасники якої мають однакову кваліфікацію.

Результати оцінювання якості UX інтерфейсу сайтів ЗВО, отримані від моделі GPT-4o та в результаті проведення колективної експертизи представлено у Таблиці 2.

Таблиця 2 - Оцінювання якості користувацьких інтерфейсів сайтів ЗВО

№	Назва ЗВО (скорочена), URL сайту	Критерій 1		Критерій 2		Критерій 3		Критерій 4	
		GPT-4o	Експерт	GPT-4o	Експерт	GPT-4o	Експерт	GPT-4o	Експерт
1	КНУ ім. Т.Шевченка www.knu.ua	3	4	2	4	2	4	3	4
2	Львівська політехніка lpnu.ua	4	4	3	4	4	4	4	5
3	МАУП www.maup.com.ua	3	2	2	2	3	2	3	3
4	КПІ ім. І.Сікорського kpi.ua	3	5	2	4	3	5	3	5
5	Запоріжська політехніка zp.edu.ua	3	3	2	2	3	3	3	3
6	ХАІ khai.edu/ua	4	4	3	4	4	4	3	5
7	ЧДТУ chdtu.edu.ua	3	2	2	3	2	2	3	3
8	УжНУ www.uzhnu.edu.ua	4	4	3	3	3	3	4	4
9	Житомирська політехніка ztu.edu.ua	4	4	3	3	5	4	4	4
10	ДНУ ім. О.Гончара www.dnu.dp.ua	3	4	2	4	2	3	3	4
11	ДУІТ duikt.edu.ua	4	4	3	4	3	4	4	4
12	ЦНТУ kntu.kr.ua	3	4	2	3	3	3	3	3
13	Одеська політехніка op.edu.ua	3	4	2	3	3	4	3	4
14	ЦДУ ім. В.Винниченка www.cusu.edu.ua	4	4	3	4	4	4	4	4
15	ЧНУ ім.Хмельницького www.cdu.edu.ua	3	2	3	4	2	2	3	3
16	ЗУНУ www.wunu.edu.ua	3	2	2	3	4	3	3	3
17	Вінницька політехніка vntu.edu.ua	3	3	2	3	3	4	3	3
18	ТНТУ ім. І.Пулюя tntu.edu.ua	3	2	3	4	2	3	3	4
19	ЧНУ ім. Ю.Федьковича www.chnu.edu.ua	4	4	3	3	5	4	4	3
20	ЛНТУ lntu.edu.ua/uk	3	4	4	4	4	5	3	3

Джерело: розроблено за результатами проведеного експертного дослідження

Для аналізу ефективності використання моделі GPT-4o в якості цифрового експерта у процесах оцінювання якості UX інтерфейсу було проведено дослідження рівня узгодженості оцінок, отриманих від GPT-4o та колективного експерта (збиральний образ оцінювача-людини) з використанням методів статистичного аналізу [18]. Для оцінки рівня узгодженості думок експертів доцільним є обчислення коефіцієнта парної рангової кореляції між бальними оцінками двох експертів (α, β) та інформаційної міри збігу експертних думок Для розрахунку коефіцієнта парної рангової кореляції використовувалась формула:

$$\rho_{\alpha,\beta} = 1 - \frac{\sum_{i=1}^{n} \varphi_i^{\,2}}{(n^3 - n)/6 + (T_1 + T_2)/12}, \quad \text{n=20}, \tag{3}$$

де φ_i - різниця (за модулем) величин рангів оцінок і-го об'єкта, заданих експертами α, β, $\varphi_i = \left| R_{i,\alpha} - R_{i,\beta} \right|$; T_1, T_2 – показники однакових рангів оцінок експертів.

Інформаційна міра збігу думок (Устюжанінова) обчислювалась за формулою:

$$I_{\alpha,\beta} = 1 - \frac{2n_{\alpha,\beta}}{n_\alpha \times \log_2(1 + n_\beta / n_a) + n_\beta \times \log_2(1 + n_\alpha / n_\beta)}, \tag{4}$$

де $n_{\alpha,\beta}$ – кількість об'єктів однаково оцінених експертами α, β; n_α, n_β - кількість об'єктів, оцінених відповідно експертом α та експертом β.

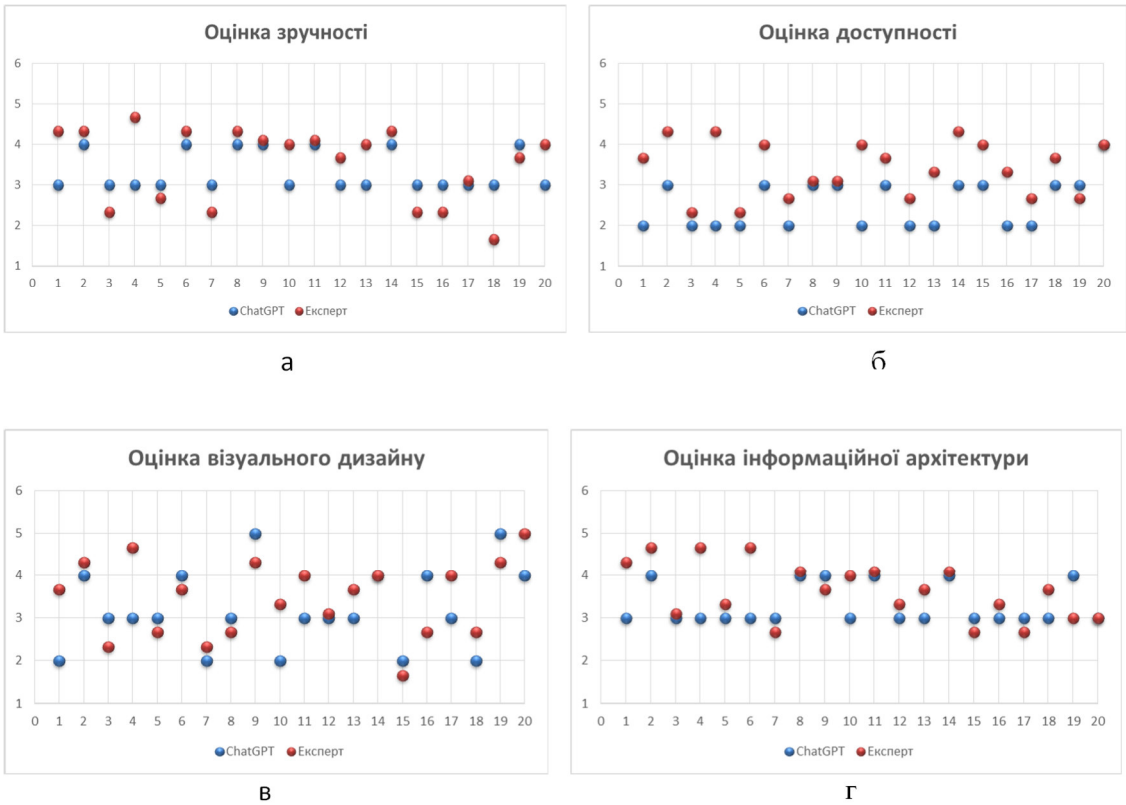
Коефіцієнти парної рангової кореляції та інформаційної міри збігу думок GPT-4o та експертів-людей обчислювались в Excel для кожного критерію окремо, результати обчислень представлено у Таблиці 3.

Таблиця 3. Результати визначення узгодженості думок експертів

Показник	Критерій 1	Критерій 2	Критерій 3	Критерій 4
$\rho_{\alpha,\beta}$	0,91	0,82	0,86	0,95
$I_{\alpha,\beta}$	0,81	0,67	0,72	1,2

Джерело: розроблено авторами за результатами проведених обчислень

Коефіцієнт попарної рангової кореляції змінюється в межах від -1 до +1. Думки експертів вважаються узгодженими, якщо коефіцієнт парної рангової кореляції перевищує 0,7. У випадках, коли $\rho \geq 0{,}9$, думки експертів вважаються сильно узгодженим. З таблиці 3 видно, що найкраще узгоджуються оцінки, отримані від GPT-4o з оцінками від експертів за критерієм 1 («зручність») і критерієм 4 («інформаційна архітектура»). На порядок нижче узгоджені оцінки GPT-4o та експертів за критерієм 3 («візуальний дизайн») і найменша міра узгодженості спостерігається за критерієм 2 («доступність»). На рис. 2 представлена візуалізація співставлення оцінок, отриманих від GPT-4o та від експертів-оцінювачів.



а – критерій 1 («зручність»); б - критерій 2 («доступність»); в - критерій 3 (візуальний дизайн»); г - критерій 4 («інформаційна архітектура»)

Рисунок 2 – Порівняльні діаграми оцінок, отриманих від GPT-4o та експертів-оцінювачів за критеріями 1-4

Джерело: розроблено авторами

Суттєва розбіжність оцінок, наданих GPT-4o та експертами-оцінювачами за критерієм 2 (рис.2 б) на нашу думку обумовлюється тим, що, як розробники так і тестувальники мають значний досвід розробки та оцінювання програмного забезпечення, але при цьому – обмежений досвід, пов'язаний із забезпеченням доступності системи, яка фактично має бути інклюзивною. Складність для розробників враховувати вимоги доступності, а тестувальникам – враховувати широкий спектр вразливостей для людей з особливими потребами також підтверджується дослідженням [13].

На другому етапі експерименту порівнювались аналітичні звіти про якість інтерфейсів, зроблені GPT-4o та експертами-оцінювачами. Аналіз відбувався за напрямками: виявлення та опис проблем в рамках одного і того ж критерію; фіксація безпекових ризиків; рекомендації по покращенню якості інтерфейсу.

Суттєвих відмінностей у показниках виявлення проблем якості користувацького інтерфейсу за критеріями 1, 3 та 4 не виявлено (рис.2 а, в, г). Але, при порівнянні аналітичних звітів про якість інтерфейсу за критерієм 2 («доступність») виявилось, що модель GPT-4o дозволяє враховувати ширший спектр вразливостей ніж той, що надається експертами. Зважаючи на більш строге оцінювання критерію 2, надане GPT-4o (що підтверджується рекомендаціями в аналітичних звітах) з метою забезпечення розробниками та тестувальниками відповідності вимогам до доступності продуктів та послуг ІКТ згідно з європейським стандартом EN 301 549 [19], вважаємо доцільним залучення інструментів на базі ШІ до оцінювання критерію доступності.

Слід також зазначити, що другий етап експерименту в нашому дослідженні грав роль допоміжного заходу для перевірки обґрунтованості отриманих бальних оцінок якості користувацького інтерфейсу (чи використовувались при цьому подібні міркування у експертів та GPT-4o) і інтерпретація отриманих результатів аналітичних звітів потребує окремого дослідження.

Висновки. У роботі представлено результати експерименту, спрямовано на вивчення ефективності GPT-4o, як інструмента для оцінювання якості користувацького інтерфейсу з урахуванням безпекових ризиків.

Основні результати дослідження:

1. Встановлено перелік критичних безпекових ризиків, притаманних обраній системі критеріїв оцінювання користувацьких інтерфейсів, а також визначено їх вплив на якість та об'єктивність UX-аналізу.

2. Розроблено двоетапну процедуру проведення експериментального дослідження.

3. Доведено статистично значущу узгодженість між оцінками GPT-4o та експертів-людей, що свідчить про можливість використання моделі як надійного суб'єкта оцінювання в процесах UX-аудиту.

4. Виявлено, що модель GPT-4o здатна виявляти UX-вразливості, зокрема щодо доступності та інклюзивності, які залишаються поза увагою експертів, що є важливим для оцінювання вебресурсів, орієнтованих на користувачів з особливими потребами.

Дане дослідження вносить певний вклад в сегмент задач автоматизації оцінювання UX-інтерфейсу, використовуючи сучасні інструменти ШІ, базові UX-показники якості інтерфейсу та пов'язані із ними безпекові ризики. Отримані результати обґрунтовують доцільність інтеграції сучасних мультимодальних моделей ШІ у процеси аудиту інтерфейсів як інструментів підсилення експертного аналізу, здатних зменшувати трудомісткість та ресурсні витрати при оцінюванні UX-інтерфейсу ІТ-систем.

Перспективи подальших досліджень включають розширення експериментальної частини із залученням експертів, як із кола тестувальників, так і фахівців із кібербезпеки, а також збільшенням кількості критеріїв для оцінки якості інтерфейсу. Крім того, не зважаючи на популярність GPT-4o, в майбутніх дослідженнях можуть бути розглянуті інші LLM, які можуть відрізнятись по згенерованим результатам та необхідним ресурсам.

Список літератури

1. Takaffoli M., Li S., Mäkelä V. Generative AI in User Experience Design and Research: How Do UX Practitioners, Teams, and Companies Use GenAI in Industry? *Designing Interactive Systems Conference*: Proc. of the ACM Int. Conf., 1-5 July 2024. Copenhagen, Denmark 2024. P. 1579–1593. DOI: 10.1145/3643834.3660720.
2. Muratovic F., Kearns-Manolatos D., Alibage A. Generative AI in Software Development: Challenges, Opportunities, and New Paradigms for Quality Assurance. *Computer*. 2025. Vol. 58, № 7. P. 31-39. DOI: 10.1109/MC.2025.3556330.
3. Wang J., Huang Y., Chen C., Liu Z., Wang S., Wang Q. Software Testing With Large Language Models: Survey, Landscape, and Vision. *IEEE Transactions on Software Engineering*. 2023. Vol. 50. P. 911-936. DOI: 10.1109/TSE.2024.3368208.
4. Ворочек О.Г., Соловей І.В. Дослідження засобів штучного інтелекту для автоматизації процесу тестування програмного забезпечення. *Вісник Національного технічного університету «ХПІ». Серія: Системний аналіз, управління та інформаційні технології*. 2024. № 1 (11)'2024. С. 58-64. DOI: 10.20998/2079-0023.2024.01.09.
5. Hsueh N.-L., Lin H.-J., Lai L.-C. Applying Large Language Model to User Experience Testing. *Electronics*. 2024. T.13, № 23. DOI: 10.3390/electronics13234633.
6. Freeman L., Robert J., Wojton H. The Impact of Generative AI on Test & Evaluation: Challenges and Opportunities. *Foundations of Software Engineering:* Proc. of the 33rd ACM Int. Conf. 23-28 June 2025. Norway. P. 1376-1380. DOI: 10.1145/3696630.3728723.
7. Quinlan M., Ceross A., Simpson A. The aesthetics of cyber security: How dousers perceive them? *arXiv preprint.* arXiv: 2306.08171vl. 2023. DOI: 10.48550/arXiv.2306.08171.
8. Petelka J., Zou Ye., Schaub F. Put Your Warning Where Your Link Is: Improving and Evaluating Email Phishing Warnings. *Human Factors in Computing Systems (CHI):* Proc. Conf. 4-9 May 2019. Glasgow Scotland Uk. No. 518. P.1-15. DOI: https://doi.org/10.1145/3290605.3300748.
9. Moran K. The Aesthetic-Usability Effect. *Nielsen Norman Group: UX-Training, Consulting, & Research:* вебсайт URL: https://www.nngroup.com/articles/aesthetic-usability-effect/ (дата звернення: 26.10.2025).
10. Duan P., Chen C.-Y., Li G., Hartmann B., Li Ya. Generating Automatic Feedback on UI Mockups with Large Language Models. *Human Factors in Computing Systems*: Proc. of the CHI Int. Conf, 11-16 May 2024. NU,United States. P.1-20. DOI: 10.1145/3613904.3642782.
11. Guerino G., Rodrigues L., Capeleti B., Mello R.F., Freire A., Zaina L. Can GPT-4o Evaluate Usability Like Human Experts? A Comparative Study on Issue Identification in Heuristic Evaluation. *arXiv preprint.* arXiv: 2506.16345v1. 2025 (дата звернення: 26.10.2025).
12. Zhong R., Hsieh G., Mcdonald D. How can LLMs support UX Practitioners with image-related tasks? *GenAICHI:* Workshop on Generative AI and HCI. 2024. P.1-6. URL: https://generativeaiandhci.github.io/papers/2024/genaichi2024_29.pdf (дата звернення: 26.10.2025).
13. Renaud K., Coles-Kemp L. Accessible and Inclusive Cyber Security: A Nuanced and Complex Challenge. *SN Computer Science*. 2022. Vol.3. DOI: 10.1007/s42979-022-01239-1.
14. Renaud K. Accessible cyber security: the next frontier? *Information Systems Security and Privacy:* Proc. of the 7th Int. Conf. 11-13 February 2021. P. 9–18. DOI: 10.5220/0010419500090018.
15. Web Accessibility in Mind: вебсайт URL: https://webaim.org/projects/million/ (дата звернення: 26.10.2025).
16. Web Content Accessibility Guidelines (WCAG) 2.1. W3C Recommendation 06 May 2025: вебсайт URL: https://www.w3.org/TR/WCAG21/?utm_source=chatgpt.com (дата звернення: 11.10.2025)
17. Flutter documentation: вебсайт URL: https://flutter.dev/docs (дата звернення: 20.10.2025).
18. Гнатієнко Г.М., Снитюк В.Є. Експертні технології прийняття рішень: Монографія. Київ: ТОВ «Маклаут», 2008, 444 с.URI: http://dspace.nbuv.gov.ua/handle/123456789/56847 (дата звернення: 28.09.2025).
19. Accessibility requirements for ICT products and services. *Harmonised European Standart*: вебсайт URL: https://www.etsi.org/deliver/etsi_en/301500_301599/301549/03.02.01_60/en_301549v030201p.pdf.

References

1. Takaffoli, M., Li, S., & Mäkelä, V. (2024). *Generative AI in User Experience Design and Research: How Do UX Practitioners, Teams, and Companies Use GenAI in Industry?* Designing Interactive Systems Conference: Proceedings of the ACM International Conference (pp. 1579–1593), July 1–5, 2024, Copenhagen, Denmark. https://doi.org/10.1145/3643834.3660720.
2. Muratovic, F., Kearns-Manolatos, D., & Alibage, A. (2025). *Generative AI in Software Development: Challenges, Opportunities, and New Paradigms for Quality Assurance*. Computer, 58(7), 31–39. doi.org/10.1109/MC.2025.3556330.
3. Wang, J., Huang, Y., Chen, C., Liu, Z., Wang, S., & Wang, Q. (2023). *Software Testing With Large Language Models: Survey, Landscape, and Vision*. IEEE Transactions on Software Engineering, 50, 911–936. https://doi.org/10.1109/TSE.2024.3368208.
4. Vorochek, O. H., & Solovei, I. V. (2024). *Doslidzhennia zasobiv shtuchnoho intelektu dlia avtomatyzatsii protsesu testuvannia prohramnoho zabezpechennia* [Research of artificial intelligence tools for automating software testing processes]. *Visnyk Natsionalnoho tekhnichnoho universytetu «KhPI». Seriia: Systemnyi analiz, upravlinnia ta informatsiini tekhnolohii*, 1(11)'2024, 58–64. https://doi.org/10.20998/2079-0023.2024.01.09 [in Ukrainian].
5. Hsueh, N.-L., Lin, H.-J., & Lai, L.-C. (2024). *Applying Large Language Model to User Experience Testing*. Electronics, 13(23). https://doi.org/10.3390/electronics13234633.

6. Freeman, L., Robert, J., & Wojton, H. (2025). *The Impact of Generative AI on Test & Evaluation: Challenges and Opportunities*. Foundations of Software Engineering: Proceedings of the 33rd ACM International Conference (pp. 1376–1380), June 23–28, 2025, Norway. https://doi.org/10.1145/3696630.3728723.
7. Quinlan, M., Ceross, A., & Simpson, A. (2023). *The aesthetics of cyber security: How do users perceive them?* arXiv preprint, arXiv:2306.08171v1. https://doi.org/10.48550/arXiv.2306.08171.
8. Petelka, J., Zou, Y., & Schaub, F. (2019). *Put Your Warning Where Your Link Is: Improving and Evaluating Email Phishing Warnings*. Human Factors in Computing Systems (CHI): Proceedings of the Conference, May 4–9, 2019, Glasgow, Scotland, UK (No. 518, pp. 1–15). https://doi.org/10.1145/3290605.3300748.
9. Moran, K. (2025, October 26). *The Aesthetic-Usability Effect*. Nielsen Norman Group: UX-Training, Consulting, & Research. https://www.nngroup.com/articles/aesthetic-usability-effect/.
10. Duan, P., Chen, C.-Y., Li, G., Hartmann, B., & Li, Y. (2024). *Generating Automatic Feedback on UI Mockups with Large Language Models*. Human Factors in Computing Systems: Proceedings of the CHI International Conference, May 11–16, 2024, NU, United States (pp. 1–20). https://doi.org/10.1145/3613904.3642782.
11. Guerino, G., Rodrigues, L., Capeleti, B., Mello, R. F., Freire, A., & Zaina, L. (2025). *Can GPT-4o Evaluate Usability Like Human Experts? A Comparative Study on Issue Identification in Heuristic Evaluation*. arXiv preprint, arXiv:2506.16345v1.
12. Zhong, R., Hsieh, G., & McDonald, D. (2024). *How can LLMs support UX Practitioners with image-related tasks?* GenAICHI: Workshop on Generative AI and HCI (pp. 1–6). generativeaiandhci.github.io/papers/2024/genaichi2024_29.pdf.
13. Renaud, K., & Coles-Kemp, L. (2022). *Accessible and Inclusive Cyber Security: A Nuanced and Complex Challenge*. SN Computer Science, 3. https://doi.org/10.1007/s42979-022-01239-1.
14. Renaud, K. (2021). *Accessible cyber security: the next frontier?* Information Systems Security and Privacy: Proceedings of the 7th International Conference, February 11–13, 2021 (pp. 9–18). doi.org/10.5220/0010419500090018.
15. Web Accessibility in Mind. (2025, October 26). *WebAIM: Web Accessibility Evaluation Report*. https://webaim.org/projects/million/.
16. World Wide Web Consortium. (2025, May 6). *Web Content Accessibility Guidelines (WCAG) 2.1*. https://www.w3.org/TR/WCAG21/?utm_source=chatgpt.com.
17. Flutter. (2025, October 20). *Flutter Documentation*. https://flutter.dev/docs.
18. Hnatiienko, H. M., & Snytiuk, V. Ye. (2008). *Ekspertni tekhnolohii pryiniattia rishen: Monohrafiia* [Expert decision-making technologies: Monograph]. Kyiv: TOV «Maklaut», 444 p. http://dspace.nbuv.gov.ua/handle/123456789/56847 [in Ukrainian].
19. European Telecommunications Standards Institute. (n.d.). *Accessibility requirements for ICT products and services*. https://www.etsi.org/deliver/etsi_en/301500_301599/301549/03.02.01_60/en_301549v030201p.pdf.

Olena Prysiazhniuk, Assoc. Prof., PhD tech. sci., **Anna Puzikova**, Assoc. Prof., PhD phi- math. sci., **Dmytro Oryshechko**
Volodymyr Vynnychenko Central Ukrainian State University, Kropyvnytskyi, Ukraine

Experimental Study of the Effectiveness of the GPT-4o Model for Evaluating the Quality of User Interfaces with Consideration of Security Risks

The emergence of multimodal large language models (LLMs), including GPT-4o, creates new opportunities for automating the evaluation of user interface quality in terms of usability, accessibility, and security. A review of publicly available sources demonstrates that the integration of LLM-based tools for user interface evaluation, particularly in the context of security risk assessment, remains insufficiently explored. The purpose of this article is to examine the effectiveness of employing the multimodal GPT-4o model for automating the assessment of web interface quality from a user experience perspective, with consideration of security risks.

The research involved the development and implementation of a set of methodological procedures, which included: 1) defining a system of evaluation criteria (usability, accessibility, visual design, and information architecture) and analyzing the associated security risks; 2) conducting a series of expert assessments of the interfaces of 20 university websites, performed by human experts and by GPT-4o using a unified set of criteria and scoring scales. In addition to numerical assessments, the experts prepared analytical reports documenting security risks identified in cases where an interface received a low score for at least one criterion, together with recommendations for improvement; 3) performing a comparative analysis of the obtained results using agreement coefficients to evaluate the consistency between GPT-4o and human evaluators.

The effectiveness of GPT-4o integration was assessed based on the degree of alignment between model-generated and human-generated scores, as well as the structure and quality of the analytical reports produced by GPT-4o. Main results of the study: 1) a set of critical security risks inherent to the selected system of user interface evaluation criteria has been identified, and their impact on the quality and objectivity of UX analysis has been determined; 2) a two-stage procedure for conducting the experimental study has been developed; 3) statistically significant agreement between the evaluations provided by GPT-4o and human experts has been demonstrated, indicating the feasibility of employing the model as a reliable assessment agent within UX audit processes; 4) it has been established that GPT-4o is capable of detecting a broader spectrum of UX vulnerabilities – particularly those related to accessibility and inclusivity – than human experts, which is essential for evaluating web resources intended for users with special needs.

interface requirements analysis, user interface quality, artificial intelligence, GPT-4o, expert evaluations, security risks

Одержано (Received) 29.10.2025 *Прорецензовано (Reviewed) 13.11.2025*
Прийнято до друку (Approved) 25.11.2025