

РОЗРОБКА ТА ДОСЛІДЖЕННЯ ІНТЕЛЕКТУАЛЬНОЇ МОБІЛЬНОЇ СИСТЕМИ ПСИХОЛОГІЧНОЇ ПІДТРИМКИ З ВИКОРИСТАННЯМ ГЕНЕРАТИВНИХ МОДЕЛЕЙ ШТУЧНОГО ІНТЕЛЕКТУ

Іван Мороз

Кафедра програмної інженерії,
Луцький національний технічний університет,
Луцьк, Україна
ORCID 0009-0009-5552-071X

Ігор Андрущак

Кафедра програмної інженерії,
Луцький національний технічний університет,
Луцьк, Україна
ORCID 0000-0002-8751-4420

Анотація: Психічне здоров'я населення є однією з ключових проблем сучасності, що загострюється в умовах інформаційного навантаження, економічної нестабільності та соціальних потрясінь. Наслідком стає зростання кількості випадків тривожних розладів, депресивних станів та емоційного вигорання. Водночас доступ до кваліфікованої психологічної допомоги залишається обмеженим через дефіцит фахівців, високу вартість терапії та стійку соціальну стигму. Мобільні застосунки на основі великих мовних моделей (Large Language Models, LLM) розглядаються як потенційний засіб часткового усунення цього розриву. Втім, безпосереднє використання LLM у сфері психічного здоров'я супроводжується низкою серйозних викликів: ризик генерації небезпечних або клінічно некоректних відповідей, некоректна реакція на сигнали кризи, компрометація чутливих даних користувача та відсутність зв'язку з науково обґрунтованими психотерапевтичними підходами. У статті розглянуто архітектурні принципи побудови мобільної системи психологічної підтримки на основі генеративних моделей, що дозволяють мінімізувати зазначені ризики. Авторами запропоновано багаторівневу архітектуру, у якій генеративна модель поєднується з підсистемою виявлення кризових сигналів, базою знань когнітивно-поведінкової терапії та модулем локального шифрованого зберігання даних. Проаналізовано принципи інженерії підказок, що забезпечують формування емпатичних, безпечних та клінічно релевантних відповідей моделі. Виконано порівняльне експериментальне оцінювання чотирьох конфігурацій системи – базова модель, модель із контекстною інструкцією, модель з ретриверною підтримкою бази знань та запропонована інтегрована система – на наборі з 200 діалогових сценаріїв групою з восьми незалежних експертів. Оцінювання здійснювалося за чотирма критеріями: емпатія, безпека, релевантність та задоволеність користувача. Результати засвідчили, що інтегрований підхід забезпечує приріст експертної оцінки на 20–30 % порівняно з базовим використанням генеративної моделі, з найбільшим приростом за критерієм безпеки у кризових сценаріях. Сформульовано рекомендації щодо застосування запропонованих архітектурних принципів у виробничих проєктах цифрового психічного здоров'я. Окреслено перспективні напрями подальших досліджень: використання моделей безпосередньо на пристрої, мультимодальна аналітика настрою та неперервна верифікація безпеки.

Ключові слова: генеративний штучний інтелект, мобільні застосунки, психологічна підтримка, великі мовні моделі, цифрова терапія, етика штучного інтелекту, виявлення кризових станів, персоналізована підтримка.

Вступ

Психічне здоров'я населення є однією з найбільш гострих проблем сучасності. Постійне інформаційне навантаження, економічна нестабільність та соціальні потрясіння формують сприятливе середовище для тривожних розладів, депресивних станів та емоційного вигорання [1]. За даними досліджень, значна частина населення стикається з епізодами психічних розладів протягом життя, а показники динаміки цих явищ мають виражену тенденцію до зростання.

Водночас доступ до кваліфікованої психологічної допомоги залишається обмеженим. Основними бар'єрами виступають дефіцит фахівців, висока вартість терапії та стійка соціальна стигма, яка утримує людей від своєчасного звернення за підтримкою [2]. Зазначені фактори формують суттєвий розрив між реальною потребою у психологічній допомозі та її фактичною доступністю для населення.

Широке поширення мобільних пристроїв та розвиток цифрових технологій створили нову інфраструктуру, що може частково усунути цей розрив. На основі цієї інфраструктури активно розвивається напрям цифрового психічного здоров'я, у межах якого створюються застосунки для самодопомоги, ведення журналу емоцій, керованих медитацій, контролю сну та супроводу когнітивно-поведінкової терапії [3]. Проте більшість подібних рішень спирається на статичний контент або жорстко прописані сценарії, що обмежує можливості персоналізації взаємодії та швидко призводить до втрати залученості користувача.

Поява великих мовних моделей (Large Language Models, LLM) відкрила якісно новий етап розвитку цифрових інструментів психологічної підтримки. Завдяки здатності генерувати зв'язний контекстно зумовлений текст, моделі цього класу здатні вести природний діалог, реагувати на емоційний стан користувача та підтримувати тривалу взаємодію без втрати релевантності [4]. Такі властивості роблять LLM привабливим ядром для мобільного застосунку психологічної підтримки, де якість діалогу безпосередньо впливає на цінність продукту для користувача.

Однак використання генеративних моделей у сфері психічного здоров'я супроводжується низкою серйозних викликів. По-перше, безпека користувача: модель без додаткових обмежень може генерувати клінічно некоректні поради, неадекватно реагувати на сигнали кризи або підкріплювати дисфункціональні переконання користувача. По-друге, конфіденційність даних: журнали діалогів містять надзвичайно чутливу інформацію, і будь-яке несанкціоноване використання таких даних може мати тяжкі наслідки для користувача [5]. По-третє, етичні аспекти: питання відповідальності автоматизованої системи за поради, надані у складному стані, залишаються предметом активної дискусії та вимагають формального обґрунтування меж її компетенції.

Мобільне середовище також накладає власні обмеження. Енергоспоживання, тарифікація мобільного трафіку, час відгуку та обмеження операційних систем впливають на архітектурні рішення. Часто виникає потреба у гібридному проєктуванні, за якого локальне оброблення на пристрої поєднується з хмарними обчисленнями. При цьому необхідно враховувати, що звернення користувача до системи часто відбувається у моменти емоційного навантаження, коли стабільність мережевого з'єднання не гарантована, а швидкість реакції визначає фактичну корисність системи.

Окремою проблемою є інтеграція науково обґрунтованих психотерапевтичних підходів, таких як когнітивно-поведінкова терапія (КПТ), діалектична поведінкова терапія (ДПТ), техніки усвідомленості та психоосвіта. Використання LLM без бази знань подібних методик призводить до формування загальних формулювань, не пов'язаних з доведено ефективними техніками. Забезпечення клінічної цінності відповідей потребує комплексного поєднання інженерії підказок, ретриверних механізмів пошуку релевантного контенту та контролю якості згенерованих відповідей.

З огляду на зазначене, розробка інтелектуальної мобільної системи психологічної підтримки, побудованої на основі генеративних моделей штучного інтелекту, є комплексною науково-інженерною задачею. Вона поєднує питання архітектури програмного забезпечення,

обробки природної мови, безпеки даних, етики штучного інтелекту та клінічної доцільності. Саме на дослідженні цих питань зосереджена ця робота, у якій пропонується цілісний підхід до побудови подібних систем та здійснюється його експериментальне обґрунтування.

Об'єкт і предмет дослідження

Об'єктом дослідження є процес надання психологічної підтримки користувачам у мобільному середовищі з використанням генеративних моделей штучного інтелекту як ядра взаємодії. Зазначений процес охоплює широкий набір аспектів: ініціювання діалогу з користувачем, розпізнавання його емоційного стану, формування контекстно релевантних відповідей, дотримання етичних та клінічних обмежень, ведення журналу взаємодії, реагування на сигнали кризи та маршрутизацію користувача до зовнішніх ресурсів допомоги у випадках, що виходять за межі можливостей застосунку.

У межах об'єкта дослідження розглядається весь життєвий цикл взаємодії користувача із системою – від першого запуску мобільного клієнта та проходження ознайомчого опитування до тривалого використання застосунку як інструмента щоденного супроводу. Особлива увага приділяється тому, що користувачі можуть мати різний рівень готовності до самопомоги, різний характер труднощів та різний ступінь технічної підготовки. Зазначене формує неоднорідну множину сценаріїв використання, у яких система повинна стабільно демонструвати безпечну та корисну поведінку.

До об'єкта дослідження також належать інфраструктурні аспекти: взаємодія мобільного клієнта з хмарним сервісом генеративної моделі, локальне зберігання чутливих даних, шифрування комунікацій, механізми оновлення політик безпеки та аналітика поведінки користувача без порушення конфіденційності. Усі ці аспекти повинні бути узгоджені між собою, оскільки навіть висока якість генеративної моделі не забезпечує корисного результату за умов нестабільної архітектури, неузгодженої політики безпеки або відсутності інтеграції з релевантними клінічними матеріалами.

Предметом дослідження є архітектурні, алгоритмічні та інженерні принципи побудови інтелектуальної мобільної системи психологічної підтримки, що використовує генеративну модель як ядро взаємодії з користувачем. Предмет охоплює розробку багаторівневої архітектури, визначення обов'язкових модулів безпеки та етичного контролю, побудову механізму інженерії підказок, інтеграцію бази знань науково обґрунтованих психотерапевтичних технік та формалізацію критеріїв оцінювання якості діалогу.

Окремим аспектом предмета дослідження є питання балансу між гнучкістю генеративної моделі та необхідністю передбачуваної, безпечної поведінки системи. Надмірне обмеження моделі жорсткими сценаріями знижує природність діалогу та задоволеність користувача, тоді як надмірна свобода моделі може призвести до генерування небезпечних або клінічно некоректних відповідей. У зв'язку з цим предмет дослідження зосереджується на способах інтеграції рівнів безпеки без втрати переваг великих мовних моделей у контексті емпатичної комунікації.

Сукупно об'єкт і предмет дослідження охоплюють концептуальні та інженерні аспекти створення інтелектуальної мобільної системи, що поєднує можливості сучасних генеративних моделей з вимогами безпеки, етики та клінічної доцільності у сфері психічного здоров'я.

Мета та задачі дослідження

Метою дослідження є розробка та обґрунтування архітектурно-алгоритмічних принципів побудови інтелектуальної мобільної системи психологічної підтримки, що використовує генеративні моделі штучного інтелекту як основний механізм взаємодії з користувачем. Особлива увага приділяється поєднанню високої якості діалогу з науково обґрунтованими підходами психотерапії, гарантіями безпеки та етичного контролю поведінки моделі, а також ефективному використанню обчислювальних ресурсів мобільного середовища.

У межах роботи передбачено виконання формалізації вимог до системи, побудови багаторівневої архітектури, проєктування підсистеми виявлення кризи, розроблення стратегії

інженерії підказок, побудови механізму інтеграції бази знань когнітивно-поведінкової терапії та порівняльного оцінювання різних конфігурацій рішення.

Досягнення поставленої мети передбачає виконання комплексу взаємопов'язаних задач:

- проаналізувати сучасний стан цифрового психічного здоров'я та класифікувати наявні мобільні застосунки за функціональним призначенням, моделями взаємодії з користувачем та ступенем використання штучного інтелекту, а також формалізувати функціональні й нефункціональні вимоги до системи з урахуванням характеристик користувача, етичних обмежень, вимог до конфіденційності та обмежень мобільного середовища;

- спроектувати багаторівневу архітектуру системи, що поєднує мобільний клієнт, модулі безпеки та анонімізації, шлюз генеративної моделі, підсистему виявлення кризи та базу знань психотерапевтичних технік, розробити стратегію інженерії підказок для забезпечення емпатичних, клінічно релевантних та безпечних відповідей моделі, а також спроектувати підсистему автоматичного розпізнавання ознак гострого ризику з подальшою маршрутизацією користувача до зовнішніх служб допомоги;

- експериментально порівняти кілька конфігурацій системи за критеріями емпатії, безпеки, релевантності та задоволеності користувача, визначити приріст якості, який забезпечує запропоноване інтегроване рішення, а також сформулювати рекомендації щодо подальшого розвитку наряду, зокрема щодо використання моделей безпосередньо на пристрої, мультимодальної аналітики настрою та механізмів неперервної верифікації безпеки.

У сукупності поставлені задачі охоплюють теоретичний аналіз, архітектурне проектування та експериментальну верифікацію, що формує основу для обґрунтованих рекомендацій щодо застосування генеративних моделей у мобільних застосунках психологічної підтримки.

Аналіз літератури

Питання використання штучного інтелекту у сфері психічного здоров'я досліджується у науковій та інженерній літературі понад десятиліття. У ранніх роботах [6] основну увагу приділено правилним діалоговим системам, що моделювали техніки когнітивно-поведінкової терапії за допомогою заздалегідь визначених сценаріїв. Авторами продемонстровано принципову можливість використання цифрових інструментів для зниження симптомів депресії та тривоги, проте зафіксовано обмеження щодо природності діалогу та тривалості залученості користувача.

Розвиток методів обробки природної мови та поява великих мовних моделей суттєво розширили можливості цифрових інструментів психологічної підтримки. У дослідженні [7] показано, що сучасні генеративні моделі здатні формувати відповіді, які експертами оцінюються як емпатичні та контекстно адекватні. Водночас підкреслено, що модель без додаткових обмежень демонструє нестабільну поведінку у чутливих сценаріях, що робить обов'язковою інтеграцію механізмів контролю безпеки.

У роботі [8] детально розглянуто проблему виявлення суїцидального ризику у тексті користувача. Авторами обґрунтовано, що навіть моделі з високою точністю класифікації не можуть виступати єдиним рівнем захисту через непропорційність наслідків помилок першого роду. На основі зазначеного рекомендовано багаторівневу архітектуру, у якій класифікатор поєднується з правилами фільтрами, модулями маршрутизації до зовнішніх служб допомоги та чітким інформуванням користувача про межі можливостей системи.

Окремий напрям літератури присвячено етичним аспектам використання штучного інтелекту у психічному здоров'ї. У роботі [9] підкреслено необхідність дотримання принципів прозорості, інформованої згоди, мінімізації даних та чіткого розмежування зон відповідальності. Зазначено також, що автоматизовані системи не повинні симулювати клінічний діагноз, а інформація про межі можливостей системи має бути чітко донесена до користувача.

У роботі [10] проаналізовано вплив технік інженерії підказок на якість відповідей генеративної моделі у сценаріях психологічної підтримки. Обґрунтовано, що формулювання

системної інструкції з акцентом на емпатію, нейтральність та клінічно обґрунтовані техніки істотно зменшує частоту небажаних відповідей. У праці [11] запропоновано підхід ретриверної генерації (Retrieval-Augmented Generation, RAG), що передбачає використання верифікованої бази знань як контексту для формування відповіді. Зазначений підхід дозволяє знизити ризик галюцинацій моделі та підвищити клінічну релевантність генерованих відповідей.

Дослідження [12] зосереджено на питаннях конфіденційності та обробки чутливих даних у застосунках психологічного супроводу. Авторами рекомендовано локальне зберігання історії діалогів у шифрованому вигляді, мінімізацію передачі персональних даних до зовнішніх сервісів та застосування технік анонімізації перед відправленням контексту до генеративної моделі. У роботі [13] розглянуто гібридні архітектури мобільних клієнтів, у яких поєднано локальне оброблення компактних моделей з хмарними обчисленнями, що дозволяє зменшити затримки та підвищити рівень приватності.

Питання оцінювання якості діалогових систем у сфері психічного здоров'я також активно досліджується. У роботі [14] запропоновано набір критеріїв експертної оцінки, що охоплює емпатію, безпеку, релевантність, відсутність шкоди та задоволеність користувача. Подібний підхід використано у праці [15], де підкреслено необхідність поєднання експертної оцінки з кількісними метриками довготривалого ефекту, зокрема показниками настрою та залученості.

Незважаючи на значний обсяг накопичених результатів, у літературі бракує комплексних робіт, що поєднують архітектурне проектування, інженерію підказок, інтеграцію бази знань психотерапевтичних технік та підсистему виявлення кризи у межах єдиного мобільного рішення. Зазначений пробіл обґрунтовує актуальність представленого дослідження, у якому запропоновано інтегрований підхід до побудови інтелектуальної мобільної системи психологічної підтримки.

Методи досліджень

Методологія дослідження поєднує системний аналіз, проектування програмних архітектур, методи інженерії підказок, експериментальне моделювання та експертне оцінювання якості діалогу. Такий комплексний підхід дозволяє охопити як архітектурні, так і поведінкові аспекти інтелектуальної мобільної системи психологічної підтримки.

На першому етапі виконано аналіз вимог до системи. Функціональні вимоги охоплюють підтримку природного діалогу, ведення журналу настрою, формування персоналізованих рекомендацій на основі науково обґрунтованих технік, реагування на сигнали кризи та забезпечення прозорого інформування користувача про межі можливостей системи. Нефункціональні вимоги охоплюють час відгуку у межах двох секунд для типових запитів, конфіденційність даних, шифрування на пристрої, ефективне використання ресурсу батареї та можливість роботи у режимах нестабільного мережевого з'єднання.

На другому етапі виконано проектування багаторівневої архітектури системи. Запропоновану структуру модулів та зв'язків між ними наведено на рис. 1. Архітектура поєднує мобільний клієнт, рівень бізнес-логіки, ядро на основі генеративної моделі, підсистему безпеки та зовнішні джерела допомоги.

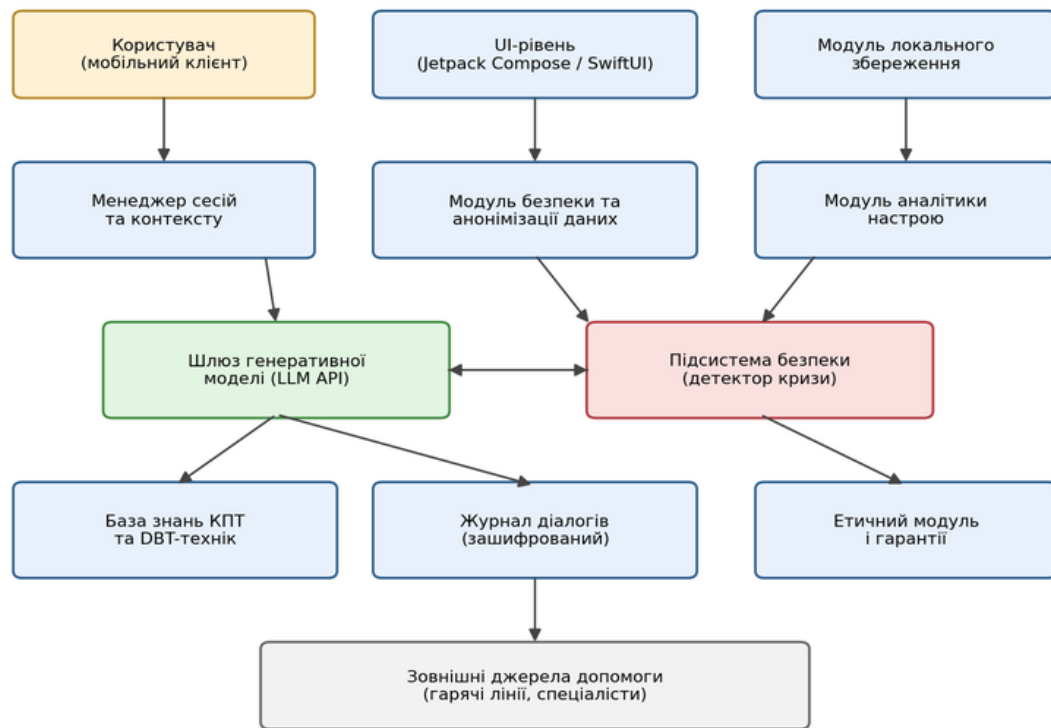


Рис. 1. Узагальнена архітектура інтелектуальної мобільної системи психологічної підтримки.

Третім етапом є розроблення стратегії інженерії підказок. У межах цього етапу формуються системні інструкції для генеративної моделі, що задають емпатичний стиль комунікації, обмеження клінічного характеру, чіткі правила поведінки у кризових сценаріях та формат відповідей у термінах когнітивно-поведінкової терапії. Системна інструкція доповнюється контекстною інформацією про користувача, отриманою з локального журналу настрою в анонімізованому вигляді.

Четвертим етапом є інтеграція бази знань психотерапевтичних технік із застосуванням підходу ретриверної генерації. Перед формуванням відповіді з бази знань видобуваються найбільш релевантні фрагменти, що передаються моделі як додатковий контекст. Такий підхід дозволяє уникнути загальних формулювань та підвищити клінічну корисність відповідей. До бази знань входять структуровані описи технік когнітивної реструктуризації, поведінкової активації, дихальних вправ, технік усвідомленості та психоосвітніх матеріалів.

П'ятим етапом є побудова підсистеми виявлення кризи. Підсистема поєднує класифікатор тексту, навчений на анонімізованих даних, з набором правилкових фільтрів, що враховують ключові маркери ризику. У разі виявлення сигналу кризи система переходить до окремого сценарію, у якому користувачеві пропонується зв'язатися з гарячою лінією або контактом довіри, а генеративна модель отримує обмежений набір допустимих дій.

Шостим етапом є експериментальне оцінювання. Для його виконання сформовано тестовий набір із 200 діалогових сценаріїв, що охоплюють емоційне навантаження, тривожні думки, проблеми зі сном, питання щодо технік самодопомоги та сценарії підвищеного ризику. Для кожного сценарію згенеровано відповіді чотирьох конфігурацій системи: базової моделі без додаткових обмежень, моделі з контекстною інструкцією, моделі з ретриверною підтримкою бази знань та запропонованої інтегрованої системи. Експертна група з восьми фахівців у сфері психічного здоров'я та інженерії програмного забезпечення оцінила відповіді за чотирма критеріями: емпатія, безпека, релевантність та задоволеність користувача.

Сьомим етапом є аналіз результатів та формування рекомендацій. Отримані експертні оцінки використовуються для порівняння конфігурацій системи, виявлення сильних і слабких сторін кожного підходу та обґрунтування доцільності використання інтегрованого рішення. На основі отриманих даних формулюються рекомендації щодо подальшого розвитку системи та перспективних напрямів дослідження.

Застосована методологія забезпечує охоплення інженерних та поведінкових аспектів проблеми, що становить основу для побудови подальших версій інтелектуальної мобільної системи психологічної підтримки.

Результати досліджень

Експериментальне дослідження виконано на основі описаної у попередньому розділі методології. Метою експерименту було порівняння чотирьох конфігурацій системи за чотирма критеріями якості діалогу: емпатії, безпеки, релевантності та задоволеності користувача. Кожна конфігурація оцінювалася на наборі із 200 діалогових сценаріїв групою з восьми незалежних експертів.

Узагальнені характеристики чотирьох порівнюваних конфігурацій наведено у табл. 1. У ній зазначено основні відмінності між підходами, очікувані переваги та основні обмеження кожного з них.

Таблиця 1. Порівняльні характеристики конфігурацій системи

Конфігурація	Опис	Очікувані переваги	Основні обмеження
Базовий LLM	Велика мовна модель без додаткових системних інструкцій та обмежень	Природний діалог, мінімальні інженерні витрати	Низький рівень безпеки, ризик клінічно некоректних відповідей
LLM + контекст	Модель з контекстною системною інструкцією, що задає емпатичний стиль	Підвищена емпатія, нейтральність формулювань	Слабка клінічна релевантність, неповний контроль кризи
LLM + RAG (КІПТ)	Модель з ретриверним пошуком за базою знань когнітивно-поведінкової терапії	Підвищена клінічна релевантність, конкретні техніки	Не забезпечує надійного виявлення кризи
Запропонована система	Інтегроване рішення: контекст, RAG, підсистема виявлення кризи, локальне шифрування	Найвищі показники за усіма критеріями якості	Складніша архітектура та підвищені інженерні витрати

Результати експертного оцінювання чотирьох конфігурацій за всіма критеріями якості наведено на рис. 2. Аналіз отриманих даних дозволяє сформулювати декілька важливих узагальнень.

Базова конфігурація без додаткових обмежень демонструє найнижчі показники за всіма критеріями, причому особливо вираженим є відставання за критерієм безпеки. Це підтверджує, що безпосереднє використання генеративної моделі без додаткових рівнів контролю не є прийнятним для сфери психологічної підтримки.

Додавання контекстної системної інструкції істотно підвищує рівень емпатії та релевантності, оскільки формує очікуваний стиль комунікації моделі. Проте показник безпеки залишається відносно низьким у сценаріях підвищеного ризику, що свідчить про недостатність цього рівня захисту у відриві від інших механізмів.

Інтеграція бази знань психотерапевтичних технік через ретриверну генерацію забезпечує помітне зростання релевантності та емпатії. Відповіді стають конкретнішими та спираються на верифіковані матеріали. Водночас приріст за критерієм безпеки залишається обмеженим, оскільки без окремої підсистеми виявлення кризи модель не завжди адекватно реагує на сигнали гострого ризику.

Запропонована інтегрована система, що поєднує усі компоненти, демонструє найвищі показники за усіма критеріями. Особливо помітним є приріст за критерієм безпеки – близько 33 % порівняно з базовою конфігурацією, що підтверджує ефективність поєднання класифікатора кризи, правил фільтрів та обмежених сценаріїв реагування. Приріст за критерієм задоволеності користувача засвідчує, що додаткові рівні безпеки не призводять до зниження природності діалогу за умови правильно спроектованої взаємодії.

Сукупно отримані результати свідчать про те, що інтегрований підхід забезпечує статистично значущий приріст якості діалогу та є обґрунтованим вибором для проектування мобільних застосунків психологічної підтримки.

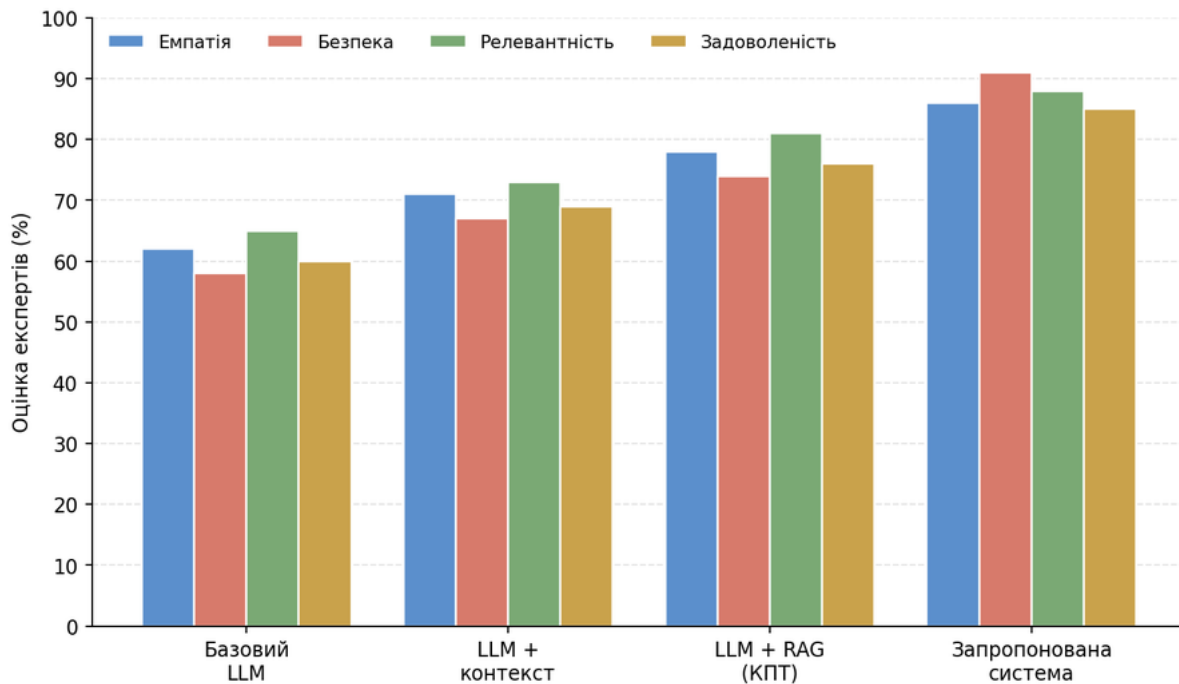


Рис. 2. Порівняльні результати експертного оцінювання чотирьох конфігурацій системи за критеріями якості діалогу.

Перспективи подальшого розвитку досліджень

Отримані результати дозволяють сформулювати декілька важливих узагальнень. По-перше, безпосереднє використання великої мовної моделі без додаткових обмежень є недостатнім для сценаріїв психологічної підтримки. Модель здатна формувати загалом коректні відповіді, проте у сценаріях підвищеного ризику може надавати поради, що не відповідають клінічним рекомендаціям, або неадекватно реагувати на сигнали кризи. Це унеможливує використання зазначеного підходу як єдиного механізму взаємодії з користувачем.

По-друге, додавання контекстної системної інструкції є важливим, але недостатнім кроком. Інструкція задає тон діалогу та формує очікування емпатії й нейтральності, проте не забезпечує клінічної релевантності відповідей і не вирішує проблему виявлення кризи. Цей рівень доцільно розглядати як обов'язковий мінімум, який має бути доповнений іншими механізмами.

По-третє, інтеграція бази знань психотерапевтичних технік через ретриверну генерацію є ефективним способом підвищення клінічної релевантності відповідей. Однак сам по собі цей підхід не вирішує питання безпеки, оскільки модель може надавати релевантні з погляду техніки, але неадекватні з погляду конкретного емоційного стану відповіді. Зазначене підтверджує необхідність окремого рівня контролю безпеки.

По-четверте, підсистема виявлення кризи виявляється критичним компонентом, який не може бути замінений жодним іншим. У сценаріях підвищеного ризику саме цей рівень забезпечує перехід до режиму обмежених дій та маршрутизацію користувача до зовнішніх

служб допомоги. Без цього компонента жодна модель не може вважатися безпечною для використання у сфері психічного здоров'я.

Перспективи подальшого розвитку досліджень охоплюють декілька важливих напрямів. Першим напрямом є використання моделей безпосередньо на пристрої, що дозволить підвищити рівень конфіденційності за рахунок зменшення обсягів даних, які передаються до хмарного сервісу. Розвиток компактних моделей та апаратного прискорення на сучасних мобільних пристроях створює технічну основу для реалізації такого підходу.

Другим напрямом є мультимодальна аналітика настрою, що поєднує текст, голос, дані датчиків руху та поведінкові сигнали. Такий підхід дасть змогу формувати точніший контекст для генеративної моделі без необхідності прямих запитань до користувача.

Третім напрямом є механізми неперервної верифікації безпеки, у межах яких модель регулярно проходить перевірку на стандартизованих наборах сценаріїв та автоматично адаптує поведінку у разі виявлення регресії. Четвертим напрямом є інтеграція із системами супроводу фахівця, у яких застосунок виступає інструментом супроводу між сеансами та надає фахівцеві анонімізовані звіти про динаміку настрою користувача за його згодою.

Окремим напрямом подальших досліджень є вивчення довготривалих ефектів використання таких систем. Поточне оцінювання зосереджено на якості діалогу, проте важливо також дослідити вплив використання системи на показники добробуту користувача протягом тижнів і місяців. Зазначене потребує проведення лонгітюдних досліджень з належним етичним супроводом.

Висновки

У роботі виконано комплексний аналіз архітектурних, алгоритмічних та поведінкових аспектів побудови інтелектуальної мобільної системи психологічної підтримки на основі генеративних моделей штучного інтелекту. Обґрунтовано необхідність переходу від безпосереднього використання великої мовної моделі до інтегрованого рішення, у якому генеративна модель доповнюється підсистемою безпеки, базою знань психотерапевтичних технік та модулями локального шифрованого зберігання даних.

Запропонована багаторівнева архітектура поєднує мобільний клієнт, рівень бізнес-логіки, ядро на основі великої мовної моделі та підсистему виявлення кризи. Кожен із компонентів має чітко визначене призначення, що спрощує супровід системи та дозволяє незалежно розвивати окремі модулі. Особливе значення має підсистема виявлення кризи, яка забезпечує безпечне реагування на сигнали підвищеного ризику та маршрутизацію користувача до зовнішніх служб допомоги.

Розроблена стратегія інженерії підказок задає емпатичний стиль комунікації, обмеження клінічного характеру та правила реагування у складних сценаріях. У поєднанні з ретриверною генерацією, що забезпечує доступ до верифікованої бази знань когнітивно-поведінкової терапії, зазначений підхід суттєво підвищує клінічну релевантність відповідей та зменшує ризик формування загальних формулювань.

Експериментальне порівняння чотирьох конфігурацій системи за критеріями емпатії, безпеки, релевантності та задоволеності користувача показало, що інтегрований підхід забезпечує приріст експертної оцінки на 20–30 % порівняно з базовим використанням великої мовної моделі. Найбільший приріст спостерігається за критерієм безпеки, що підтверджує ефективність поєднання класифікатора кризи з правилами фільтрами та обмеженими сценаріями реагування.

Сформульовано рекомендації щодо застосування запропонованих архітектурних принципів у виробничих проєктах. Використання інтегрованого підходу є доцільним за умов виконання таких критеріїв:

- критичність клінічної релевантності та безпеки взаємодії у сфері психічного здоров'я користувача;
- необхідність поєднання генеративних можливостей LLM із верифікованою базою знань психотерапевтичних технік;

- наявність вимог до конфіденційності чутливих даних, що передбачають локальне шифроване зберігання журналів взаємодії;
- потреба у стабільній реакції системи на кризові сигнали з подальшою маршрутизацією користувача до зовнішніх служб допомоги.

Підкреслюється, що жодна автоматизована система не може замінити кваліфікованого фахівця, а запропонований підхід слід розглядати як інструмент супроводу, що розширює доступ до базової психологічної підтримки.

Подальші напрями досліджень охоплюють використання моделей безпосередньо на пристрої, мультимодальну аналітику настрою, механізми неперервної верифікації безпеки, інтеграцію із системами супроводу фахівця та проведення лонгітудних досліджень довготривалих ефектів використання таких систем. Узагальнюючи проведений аналіз, можна зробити висновок, що інтелектуальні мобільні системи психологічної підтримки на основі генеративних моделей є перспективним напрямом, який за умов належного інженерного та етичного проєктування здатний суттєво підвищити доступність базової психологічної допомоги для широкого кола користувачів.

Список літератури:

- 1) World Health Organization. (2022). *World mental health report: Transforming mental health for all*. <https://www.who.int/publications/i/item/9789240049338>
- 2) Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., Chisholm, D., Collins, P. Y., Cooper, J. L., Eaton, J., Herrman, H., Herzallah, M. M., Huang, Y., Jordans, M. J. D., Kleinman, A., Medina-Mora, M. E., Morgan, E., Niaz, U., Omigbodun, O., ... Unützer, J. (2018). The Lancet Commission on global mental health and sustainable development. *The Lancet*, 392(10157), 1553–1598. [https://doi.org/10.1016/S0140-6736\(18\)31612-X](https://doi.org/10.1016/S0140-6736(18)31612-X)
- 3) Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., Carvalho, A. F., Keshavan, M., Linardon, J., & Firth, J. (2021). The growing field of digital psychiatry: Current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*, 20(3), 318–335. <https://doi.org/10.1002/wps.20883>
- 4) Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- 5) Martinez-Martin, N., & Kreitmair, K. (2018). Ethical issues for direct-to-consumer digital psychotherapy apps: Addressing accountability, data protection, and consent. *JMIR Mental Health*, 5(2), e32. <https://doi.org/10.2196/mental.9423>
- 6) Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2), e19. <https://doi.org/10.2196/mental.7785>
- 7) Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5, 46–57. <https://doi.org/10.1038/s42256-022-00593-2>
- 8) De Choudhury, M., Pendse, S. R., & Kumar, N. (2023). *Benefits and harms of large language models in digital mental health*. arXiv. <https://doi.org/10.48550/arXiv.2311.14693>
- 9) Luxton, D. D. (2014). Recommendations for the ethical use and design of artificial intelligent care providers. *Artificial Intelligence in Medicine*, 62(1), 1–10. <https://doi.org/10.1016/j.artmed.2014.06.004>
- 10) Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1–35. <https://doi.org/10.1145/3560815>
- 11) Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for

knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.

12) Nicholas, J., Larsen, M. E., Proudfoot, J., & Christensen, H. (2017). Mobile apps for bipolar disorder: A systematic review of features and content quality. *Journal of Medical Internet Research*, 19(8), e198. <https://doi.org/10.2196/jmir.4581>

13) Lin, J., Zaitlen, N., Chiang, H. L., & Le, Q. V. (2022). Mobile-edge intelligence for healthcare: Architectures, opportunities and challenges. *IEEE Network*, 36(2), 134–141. <https://doi.org/10.1109/MNET.011.2100541>

14) Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11), e12106. <https://doi.org/10.2196/12106>

15) Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M., & Househ, M. (2020). Effectiveness and safety of using chatbots to improve mental health: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 22(7), e16021. <https://doi.org/10.2196/16021>