УДК 004.8:61
DOI https://doi.org/10.32782/2663-5941/2025.5.2/15

Кандель К.В.
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

ЗАСТОСУВАННЯ ВІРТУАЛЬНИХ АСИСТЕНТІВ У МЕДИЦИНІ

У роботі узагальнено підходи до побудови та інтеграції віртуальних асистентів у медичні інформаційні системи на базі чотирьох взаємодоповнювальних архітектур. Розглянуто FHIR-центровані мікросервіси як транзакційний «каркас» обміну даними, вбудовані до електронної медичної карти Copilot-рішення для «ambient»-документації в ході прийому, хмарні великі мовні моделі з пошуком та отриманням знань (RAG) поверх локальних джерел, також доменні чат-боти для вузьких клінічних сценаріїв (тріяж, самообслуговування). Показано, що якість роботи віртуальних асистентів визначається зрілим NLP-конвеєром, вилучення клінічних фактів, нормалізація та прив'язка до термінологій (UMLS, SNOMED CT, LOINC, RxNorm) з подальшим формуванням структурованих FHIR-ресурсів та аудитом. Узагальнено наявні емпіричні дані щодо користі розмовних агентів у психічному здоров'ї (метааналізи РКД демонструють клінічно значущі ефекти), підкреслено вимоги до лікарського нагляду й безпеки. Наведено порівняльні переваги й обмеження підходів, мікросервіси забезпечують масштабованість та інтероперабельність, але ускладнюють експлуатацію, Copilot зменшує документальне навантаження, проте залежить від постачальника ЕМК та якості розпізнавання мовлення, RAG-архітектури підвищують перевірюваність відповідей за рахунок посилань на джерела, водночас потребують суворого контролю доступу й актуальності індексу, доменні боти дають передбачувану поведінку у вузьких процесах, але масштабуються ціною розмноження сценаріїв і правил. Обґрунтовано комбінований підхід, у якому FHIR-мікросервіси забезпечують обмін та цілісність даних, Copilot працює у кабінеті лікаря, RAG-шар підтримує пошук і обґрунтування, а спеціалізовані боти виконують завершені дії. Окреслено кроки локалізації для України через двомовні інтерфейси, україномовний терміношар, зіставлення довідників до SNOMED CT/LOINC, профілювання FHIR під процеси НСЗУ, вимоги юрисдикції зберігання даних та повний журнал аудиту.

Ключові слова: *мікросервіси, семантичний шар, конфіденційність даних, розпізнавання мовлення, термінологічне зіставлення, архітектура.*

Постановка проблеми. Віртуальні асистенти в медицині, це програмні агенти, що інтегруються з клінічними системами, структурують дані, автоматизують запис і документацію та підтримують рішення за стандартами. Їх підклас розмовні асистенти, також іменовані як чат-боти, це програми, призначені для генерації природного тексту відповідно до заданого контексту. Вони все частіше використовуються у різних сферах, включаючи охорону здоров'я. Забезпечуючи кращу доступність, персоналізацію та ефективність, розмовні агенти поліпшують догляд за пацієнтами, взявши на себе значну роль у сприянні здоров'ю та благополуччю користувачів. Чат-боти стали повсюдними в нашому житті, забезпечуючи природні розмови з користувачами через різні канали комунікації. У міру зростання кількості досліджень та доступних інструментів, пов'язаних з чат-ботами, виникає критична необхідність оцінювати їхні характеристики, щоб покращувати дизайн чат-ботів, які ефективно впливають на здоров'я.

Аналіз останніх досліджень і публікацій. У науковому просторі сьогодення представлено значну кількість досліджень, спрямованих на оцінку ефективності медичних чат-ботів. Так у статті [1] подано систематичний огляд розмовних агентів у сфері охорони здоров'я. Із 1513 записів автори відібрали 17 досліджень (14 різних агентів), де переважали кінцево-станова та каркасна стратегії керування діалогом, а агент-орієнтовані підходи траплялися рідко. Половина систем підтримувала самодопомогу та зміну поведінки, доказів ефективності небагато – лише два РКД і одне кросоверне, причому одне випробування показало зниження симптомів депресії. Оцінка безпеки проводилася нечасто, тож потрібні стандартизовані протоколи якості та безпеки.

Дослідження [2] розглядає ефективність ШІ-чатботів у медичній сфері у форматі систематичного огляду. До фінального аналізу включено

31 дослідження, що оцінювали агентів у профілактиці, психічному здоров'ї, менеджменті хронічних захворювань і зміні способу життя. Переважна більшість робіт повідомляє про позитивні клінічні або поведінкові ефекти та високу прийнятність для користувачів, однак доказова база обмежена невеликими вибірками, короткою тривалістю втручань і значною гетерогенністю, через що метааналіз виконати неможливо. Оцінка безпеки проводиться рідко. Автори роблять висновок про обнадійливий потенціал технології та потребу в більш суворих РКД із стандартизованими показниками і довшим спостереженням.

Дослідження [3] подає кейс-стаді реального застосування чатботів для самодіагностики у повсякденних умовах. Автори аналізують журнали взаємодій користувачів із діючим сервісом, оцінюючи профіль аудиторії, типові симптоми, тривалість і структуру діалогів, частку завершених сесій та подальші кроки користувачів. Показано стійку залученість, переважання запитів щодо поширених скарг і корисність бота для первинного скринінгу та інформаційної підтримки, зокрема поза робочим часом. Водночас відзначено обмеження: неповна клінічна валідація, потреба у зрозумілих рекомендаціях і безпечних сценаріях ескалації до лікаря, а також вимоги до захисту даних та постійного моніторингу якості сервісу.

Дослідження [4] формулює базові метрики для порівнюваного оцінювання медичних чатботів. Автори пропонують ядро показників за кількома вимірами, зокрема клінічна корисність і відповідність керівництвам, безпека й уникнення шкідливих порад з обов'язковою ескалацією невизначеності, своєчасність і доступність, зручність та утримання користувачів, а також справедливість і відсутність упереджень. Рекомендовано поєднувати рандомізовані та «in-the-wild» дослідження, стандартизувати сценарні тест-набори, фіксувати деномінатори, конфаундери та серйозні інциденти. Наголошено, що із поширенням LLM оцінка має виходити за межі суто лінгвістичних бенчмарків і відображати реальний вплив на клінічні процеси та результати лікування.

Проте, незважаючи на значну кількість наукових надбань, питання інтеграції ВА у FHIR-екосистему, термінологічної локалізації для України та проспективної оцінки в первинній ланці лишається відкритим.

Метою статті є створення науково обґрунтованої рамки впровадження віртуальних асистентів у медицині, що поєднує FHIR-центровані мікросервіси, вбудовані в ЕМК, хмарні LLM із RAG та доменні чат-боти. Завдання статті: систематизувати й порівняти ці архітектури та NLP-конвеєр (вилучення фактів, нормалізація, лінкування до SNOMED CT/LOINC/RxNorm), визначити семантичний шар (UMLS тощо), описати інтеграційні схеми й вимоги безпеки, а також запропонувати метрики оцінювання якості та план локалізації для України (профілі FHIR НСЗУ, двомовність, юрисдикційні вимоги до зберігання даних).

Виклад основного матеріалу. У недавніх рандомізованих дослідженнях чат-боти на основі ШІ демонструють клінічно значуще зниження депресивної симптоматики. Метааналіз 15 РКД (npj Digital Medicine, 2023) [5] показав статистично значуще зменшення депресії з сумарним ефектом Hedges' g ≈ 0,64 (95% ДІ 0,17–1,12), найбільші ефекти відзначалися у генеративних та мультимодальних/голосових агентів, особливо у випадку мобільних програм та месенджерів. Додатково, метааналіз 2024 (Journal of Affective Disorders) [6] підтвердив короткострокове покращення депресії та тривоги у дорослих при використанні чатботів.

Використання розмовних медичних агентів з вільним введенням вже широко застосовується у дослідженнях, їхня ефективність та безпека оцінюються за стандартними протоколами, включаючи проспективні дизайни, а для окремих завдань (в т.ч. контроль за станом пацієнтів з депресією) є рандомізовані дані про клінічно значуще покращення. У практичних сценаріях, особливо при збиранні анамнезу, такі системи підвищують залучення пацієнтів, структурують первинний збір даних та підтримують прийняття клінічних рішень.

У статті розглядається один тип генеративного ШІ – великі мультимодальні моделі (ВММ – large multimodal model – LMM), які можуть приймати один або декілька типів вхідних даних та генерувати різноманітні результати, не обмежені типом цих вхідних даних. На 2025 рік ВММ вже широко застосовуються в дослідженнях та допоміжних клінічних сценаріях (наприклад, чернетки відповідей пацієнтам та «ambient»-документація), під контролем клініцистів. Автономні діагностичні застосування залишаються предметом досліджень, їхнє впровадження потребує валідованих протоколів безпеки та відповідності керівництву ВООЗ.

Для порівняльного розгляду відібрано чотири взаємодоповнювальні архітектури віртуальних асистентів (ВА). Перша з них – FHIR (Fast Healthcare Interoperability Resources), що є мікросервісною архітектурою, що будується навколо семантичного шару – стандарту HL7 FHIR, який є еталонною моделлю даних [7]. Кожен сервіс відповідає

за окрему предметну область, тобто за пацієнтів, розклади, записи на прийом, спостереження та опитувальники, управлінням доступу та аудитом. Взаємодія між сервісами та зовнішніми системами виконується через веб-інтерфейси та обмін подіями [8]. Щоразу, коли створюється чи змінюється ресурс, наприклад, запис на прийом чи результат опитувальника, публікується подія, та інші послуги реагують у межах своєї ролі. Віртуальний асистент виступає клієнтом цієї архітектури, після розуміння наміру користувача він викликає відповідні операції, оформлюючи їх у термінах ресурсів FHIR та профілів, узгоджених з локальною інформаційною системою охорони здоров'я.

Такий підхід дає можливість вести всі обміни FHIR-ресурсами та словниками, що дозволяє під'єдуватися до електронної медичної карти (ЕМК), лабораторій та реєстрів. Декомпозиція на невеликі сервіси знижує зв'язаність, дозволяє незалежно розгортати та масштабувати компоненти, прискорює випуск нових функцій (наприклад, тріаж чи самозапис) [9].

Однак, модульність компенсується складнішою експлуатацією, тобто в результаті в структурі буде більше вузлів та точок відмови. Наскрізні процеси («записати пацієнта та відправити нагадування») вимагають узгодженої оркестрації та механізмів відновлення при часткових збоях. Потрібно вести версії профілів FHIR та відображення з успадкованих кодів, підтримувати термінологічний шар. Міграція історичних даних та тестування контрактів є трудомісткими, а інфраструктурні та спостережувані витрати вищими, ніж у монолітної архітектури.

Друга архітектура, чат Copilot вбудовується прямо в електронну медичну карту і працює паралельно з лікарським прийомом [10]. Мова консультації захоплюється через безпечний канал, далі модуль розпізнавання мови перетворює його на текст із поділом ролей лікаря та пацієнта [11]. Модуль обробки природної мови фіксує скарги, анамнез, об'єктивні дані, призначення та рекомендації, після чого генератор формує чернетку запису за локальними шаблонами. Структуровані елементи заповнюються у відповідні поля карти, при необхідності створюються записи про спостереження, призначення та направлення. Перед збереженням лікар переглядає та підтверджує чернетку [12], зміни фіксуються в журналі, джерелом вважається аудіозапис візиту.

Знижується час на оформлення документації та зменшується адміністративне навантаження, що дозволяє приділяти більше уваги пацієнтові. Підвищується повнота та однаковість записів завдяки шаблонам та автоматичному вилученню клінічних фактів. Прискорюється внесення результатів обстежень та рекомендацій, легше підтримувати вимоги контролю якості та безпеки завдяки журналу та відстеженню змін. До того ж, інтеграція в карту усуває зайві перемикання між системами.

З-поміж недоліків цієї архітектури варто виявити те, що виникає залежність від постачальника медичної інформаційної системи та її вбудованих інтерфейсів. Можливі помилки інтерпретації та неповне відображення контексту, тому потрібен обов'язковий перегляд та редагування лікарем. Закритість моделей знижує прозорість прийняття рішень, а обробка аудіо та тексту збільшує споживання обчислювальних ресурсів та висуває підвищені вимоги до захисту персональних даних [13].

Третя архітектура, а саме велика мовна модель в межах хмарної системи підключається до затвердженого корпусу клінічних даних [14]. Поверх корпоративних сховищ розгортається шар пошуку та отримання знань. Він індексує документи, записи електронної карти, результати досліджень, словники та онтології [15]. За запитом асистента або лікаря модуль вилучення формує коротку добірку релевантних фрагментів з метаданими джерел. Ця добірка разом з інструкціями щодо стилю та безпеки передається мовній моделі. Модель будує відповідь або чернетку клінічного тексту і посилається на знайдені фрагменти [16]. Для трансакційних дій використовуються профільовані інтерфейси обміну даними. Збереження виконується через узгоджені ресурси та з журналуванням. Доступ розмежовується за ролями, всі звернення фіксуються в аудиті.

Єдина точка доступу до знань знижує фрагментацію інформації та прискорює прийняття рішень. Відповіді спираються на локальні дані, а наявність посилань підвищує перевірюваність. Архітектура масштабується горизонтально і дозволяє швидко додавати нові типи джерел без змін в інтерфейсі користувача. Прискорюється підготовка виписок, напрямів та рекомендацій, спрощується пошук рідкісних протоколів та локальних регламентів. Інтеграція поверх існуючих сховищ дає ефект без глибокої обробки медичної інформаційної системи. Наявність термінологічного шару полегшує узгодження понять та зменшує кількість невідповідностей.

З негативної точки зору варто зазначити, що виникають вимоги до захисту персональної інформації та юрисдикції обробки, а також до деталь-

ного розмежування прав доступу. Вартість обчислень та затримки відповіді вища, ніж у локальних правил, що важливо при масовому навантаженні. Потрібні процедури контролю якості, моніторинг посилань та регулярна переіндексація. Складність експлуатації зростає за рахунок версіонування джерел та шаблонів підказок.

Четверта архітектура, доменний чат-бот, що будується навколо вузького медичного завдання [17]. У центрі знаходиться модуль розуміння природного мовлення. Він виділяє наміри користувача та розпізнає клінічні концепти та факти в тексті, нормалізує їх до кодів SNOMED CT/LOINC/RxNorm та представляє як FHIR-ресурси [18]. Поруч з ним працює менеджер діалогу. Політики діалогу задають логіку кроків розмови, умови переходів, психологічні тригери до людини та правила безпеки. Знання беруться із перевіреної бази. У ній лежать протоколи, локальні регламенти, словники та онтології. Чат-бот підключається до зовнішніх систем через конектори, включаючи запис на прийом, електронну медичну картку [19], лабораторію, аптеку, домашні прилади та послуги сповіщень. Є профіль пацієнта та сховище контексту. Передавання даних захищене, усі операції фіксуються в журналі [20]. Налаштування та шаблони сценаріїв керуються через панель конфігурації.

Чітка спеціалізація дає передбачувану поведінку та зрозумілі межі відповідальності. Простіше дотримуватись вимог безпеки та приватності, локальні протоколи швидко вбудовуються в логіку розмови, а конектори забезпечують завершені дії. Таким чином пацієнт може записатися, передати дані, отримати нагадування та інструкції. Підтримується робота на декількох каналах зв'язку та рідною мовою користувача. Навантаження розподіляється між невеликими компонентами, знижуючи затримки та час обчислень.

Вузька спеціалізація обмежує охоплення завдань, а при розширенні тематики зростає кількість сценаріїв, правил та тестів. Конектори вимагають підтримки версій та постійного зіставлення термінів, при цьому неузгодженість словників веде до помилок. Оновлення протоколів та керування змінами можуть стати вузьким місцем, посилюючи ризики залежності від конкретного постачальника та ускладнюючи забезпечення єдиного досвіду між каналами та пристроями. Оцінка якості вимагає доменних датасетів та регулярного аудиту, а їхня підготовка витратна. При нестачі контексту або появі нетипової ситуації потрібне своєчасне переведення розмови до фахівця.

Порівнюючи чотири архітектурні підходи, видно, що вони вирішують різні завдання і добре доповнюють одне одного. FHIR-центровані мікросервіси дають транзакційний основу системи та міжсистемну сумісність, тому найкраще підходять для запису на прийом, обміну результатами та надійного аудиту. Вбудовані в ЕМК Copilot знімають рутинне навантаження на лікаря і видають чернетки візитів, зате сильніше залежать від постачальника і завжди вимагають лікарського контролю. Хмарні моделі з отриманням знань з локального корпусу впевнено відповідають на запитання та допомагають із сумаризацією регламентів та виписок, але коштують дорожче у обчисленнях та висувають підвищені вимоги до захисту даних та розмежування доступу. Доменні чат-боти супроводжують пацієнта у вузьких сценаріях на зразок тріажу чи самообслуговування і виконують завершені дії через конектори, проте масштабуються переважно шляхом додавання нових сценаріїв і потребують суворого керування правилами. Якщо потрібен єдиний ландшафт, оптимальна комбінація, де мікросервіси на FHIR забезпечують обмін та цілісність, Copilot працюють у кабінеті лікаря, шар здобуття знань підтримує пошук та обґрунтування відповідей, а спеціалізовані боти автоматизують пацієнтські маршрути. Вибір часток кожного підходу визначається клінічною критичністю завдань, цільовими затримками відгуку, чутливістю даних та вимогами регулятора.

Алгоритми обробки природної мови (Natural Language Processing – NLP) забезпечують віртуальним помічникам розуміння намірів, заповнення структурованих полів медичної карти та формулювання рекомендацій.

До фундаментальних завдань алгоритмів NLP варто віднести: вилучення клінічних фактів з тексту, нормалізація понять та кодування в термінологіях, вилучення відносин та тимчасових зв'язків, класифікація документів та автокодування, сумаризація та перефраз, відповіді на питання та діалогове управління, управління де-ідентифікація медичного тексту, обробка мови та мультиканали.

У роботі про BioBERT (Lee J.) [21] це реалізовано як доменно-орієнтоване переносне навчання для біомедичного текст-майнінгу. Автори виходять з того, що біомедичні корпуси насичені спеціалізованими термінами та власними іменами, через що моделі загального призначення деградують, тому базовий BERT донавчається на текстах PubMed і PMC з різними комбінаціями корпусів і тривалістю попереднього навчання. Це знижує розрив

домену та покращує перенесення на завдання отримання знань з літератури. BioBERT зберігає архітектуру BERT і використовує WordPiece-токенізацію, щоб пом'якшити ефект рідкісних термінів та друкарських помилок, при цьому автори віддають перевагу «cased» словнику та сумісності з вихідним BERT для спрощення обміну вагами та повторного використання моделей. На рівні експлуатації NLP-конвеєра модель служить універсальним енкодером, поверх якого донавчаються тонкі вихідні шари для наступних завдань – розпізнавання сутностей (NER) через токен-класифікацію, вилучення відносин (RE) через класифікацію пропозиції на основі подання [CLS] з анонімізацією цільових сут SQuAD/BioASQ. Експериментально показано, що таке доменне попереднє навчання системно підвищує метрики, приріст F1 у NER та RE та MRR у QA щодо попередніх SOTA при мінімальних архітектурних змінах. Технічно описані режими попереднього навчання, обсяги даних та ресурси, що підтверджують виробничу реалізацію підходу для клінічного NLP.

У роботі (Singhal K.,) [22] Med-PaLM 2 представлений як велика мовна модель для медицини, навчена на покращеній базі PaLM 2 з доменним донавчанням та спеціальними стратегіями підказок, що переводить систему з класу енкодерів під завдання до класу генеративних моделей з медичним міркуванням. Автори відзначають три ключові елементи – сильнішу базову LLM, цільове медичне донавчання та ансамблеве уточнення як спосіб покращити ланцюжки міркувань, на цій базі отримані прирости на піднаборах MultiMedQA та суттєве покращення за MedQA/USMLE порівняно з Med-PaLM та Flan-PaLM. На відміну від BioBERT, який служить переносним енкодером з тонкими вихідними шарами для NER/RE/QA, Med-PaLM 2 генерує довгі відповіді, орієнтуючись на клінічні критерії корисності та безпеку, в парних лікарських оцінках відповіді Med-PaLM 2 частіше визнавали кращими, зокрема за критерієм неточності та нерелевантності лікарі іноді перевершували модель. До плюсів щодо BioBERT відносяться здатність давати обґрунтовані довгі відповіді, покращена медична міркування та людино-орієнтована оцінка якості, аж до переваг відповідей лікарями за більшістю критеріїв. До обмежень – неповна закритість ризиків безпеки та зсувів, необхідність подальшої валідації у реальних сценаріях, а також випадки, коли відповіді містять неточності чи надмірності та потребують нагляду. У сумі, Med-PaLM 2 – перехід від модульних енкодерних рішень біомедичного NLP до спеціалізованої медичної LLM, придатної для клінічних сценаріїв питань-відповідей за наявності процедур контролю та валідації.

В роботі (Tu T.,) [23] розглядається система AMIE, яка переводить обробку природної мови від модульних енкодерних рішень і окремих моделей запитань-відповідей до керованого діагностичного діалогу з цільовою оптимізацією під збирання анамнезу, диференціальне міркування та якість комунікації. Система донавчається на реальних та симульованих медичних діалогах, використовує двоконтурне самонавчання в симульованому середовищі з автоматизованим зворотним зв'язком та застосовує покрокове міркування для послідовного уточнення введення та відповіді у кожному ході розмови. Це виводить AMIE за рамки BioBERT, який служить переносним енкодером для NER/RE/QA, і Med-PaLM 2, орієнтованої на відповіді щодо медичних бенчмарків, але не на повний клінічний цикл – збір скарг, формування гіпотез, план ведення та пацієнт-центрична комунікація. Верифікація також інша, замість офлайнових тестів AMIE порівнюють з лікарями первинної ланки в рандомізованій стандартизованій клінічній оцінці (OSCE) у дистанційному форматі з валідованими симульованими пацієнтами та рубрикою, яка одночасно вимірює діагностичну точність, повноту збору даних, обґрунтованість плану та критерії. У цих умовах AMIE показав більш високу точність диференціальних діагнозів та перевагу за більшістю критеріїв оцінки без статистично значущого погіршення щодо інших. Сильна сторона щодо BioBERT та Med-PaLM 2 – цільова оптимізація під багатоходовий клінічний діалог та проспективна оцінка взаємодії. Обмеження – текстовий чат як інтерфейс, який відрізняється від звичайної практики, та необхідність додаткової валідації, заходів безпеки та нагляду перед впровадженням у реальні потоки.

Онтології та знання-орієнтовані системи забезпечують віртуальним асистентам однозначну інтерпретацію медичного тексту, діалогу та транзакцій, перетворюючи їх на машиночитані коди та факти для обміну та підтримки рішень.

Класифікація компонентів семантичного шару

1. UMLS – інтеграційний метатезаурус та інструменти лексичної нормалізації [24].

2. SNOMED CT – уніфікована клінічна термінологія для станів, процедур та ознак.

3. LOINC – ідентифікатори лабораторних та клінічних спостережень.

4. RxNorm – нормалізовані найменування та ідентифікатори ліків.

5. Конвеєр розпізнавання та нормалізації клінічних понять (витяг → нормалізація → прив'язка до онтологій).

6. Знаннєві графи – представлення домену вузлами та зв'язками для пошуку та виведення.

7. Правила та онтології в CDSS – формалізація клінічної логіки та її виконання.

UMLS агрегує безліч словників та кодових систем, дає єдину сітку понять та лексичні інструменти для зіставлення текстових згадок із концептами. На відміну від вузьких словників, тут покриття ширше, а вхід до NLP-конвеєра простіше, розпізнані факти можна нормалізувати в загальну систему координат і потім переводити в потрібні коди. Сильні сторони – зниження трудомісткості первинної нормалізації, можливість багатословних сценаріїв та багатомовних входів. Обмеження – необхідність налаштування під локальні корпуси, версіонування джерел та неоднорідність якості за доменом. У контурі асистента UMLS виступає як базовий шар, через який проходять результати вилучення, після чого дані готові до кодування та обміну.

SNOMED CT задає детальну клінічну ієрархію [25] для симптомів, діагнозів та процедур, LOINC фіксує спостереження та лабораторні тести [26], RxNorm нормалізує лікарські призначення [27]. У порівнянні з UMLS, який склеює словники, ці три ресурси дають стабільні цільові коди для запису в медичну картку та правил підтримки рішень. Сильні сторони – інтероперабельність на рівні архітектури, підвищення якості даних, можливість агрегувати та повторно використати факти. Недоліки – потрібні постійні процедури зіставлення локальних довідників, ведення версій та термінологічна експертиза, а за відсутності автоматизації зростають витрати та ризик невідповідностей. Для асистента це фіксовані точки прив'язки, саме ці коди входять до ресурсів обміну та ініціюють клінічні правила.

онвеєр розпізнавання та нормалізації клінічних понять (витяг → нормалізація → прив'язка до онтологій) описує шлях клінічного тексту до кодів. Спочатку з потоку мови або замітки виділяються факти – симптоми, діагнози, ліки, значення аналізів, а також тимчасові та модальні ознаки типу заперечень. Далі виконується нормалізація формулювань та величин, приведення до канонічної форми, одиниць, словникових синонімів. Потім запускається прив'язка до онтологій до термінології. Генеруються кандидати з UMLS та локальних словників, ранжуються за контекстом ембеддингами та правилами, вибирається код SNOMED CT, LOINC або RxNorm. До кожного результату додаються оцінка впевненості та відомості про джерело. На виході формуються структуровані записи у форматі FHIR та створюється журнал для аудиту. За низької впевненості результат спрямовується на перевірку фахівцем. Ведення версій словників та наборів значень здійснюється централізовано.

Знанієві графи об'єднують ЕМК, літературу та довідники [28], представляючи дані як триплети суб'єкт-предикат-об'єкт і тим самим підтримуючи складні запити та зрозумілі відповіді. До їх переваг належать гнучка інтеграція джерел та персоналізація на рівні зв'язків. До обмежень відноситься залежність якості від методики вилучення та необхідність регулярної валідації та оновлення. Правила на онтологіях OWL та SWRL та відповідні движки виконання забезпечують формальну клінічну логіку та аудит її застосування [29]. Їхні сильні сторони – прозорість і трасування. Слабкі сторони – низька переносимість правил між системами та витрати на супровід. У контексті NLP-конвеєра саме тут результати прив'язки до онтологій перетворюються на рішення, коли нормалізовані факти подаються до графа та правил і формують обґрунтовані рекомендації. Далі вони серіалізуються у стандартизовані записи для обміну та подальшого аналізу.

Віртуальні асистенти в медицині вже поєднують чотири перевірені підходи – FHIR-центровані мікросервіси для транзакцій та обміну, Copilot в ЕМК для «ambient»-документації, LLM із отриманням знань за локальними даними для відповідей та сумаризації, а також доменні чат-боти для спеціалізованих пацієнтських сценаріїв. Якість забезпечують конвеєр NLP вилучення фактів потім нормалізація та прив'язка до кодів, семантичний шар UMLS, SNOMED CT, LOINC, RxNorm, аудит та лікарський нагляд.

Для локалізації технології в Україні потрібно створити двомовні інтерфейси та повноцінний україномовний шар термінів. Зіставлення локальних довідників із SNOMED CT та LOINC та ведення актуальних наборів значень. Профілі FHIR під національні реєстри та процеси НСЗУ, підтримка е-напряму та е-рецепту, зберігання та обробка даних з урахуванням юрисдикції та журналування.

Висновки. У роботі узагальнено чотири взаємодоповнювальні архітектури віртуальних асистентів у медицині – FHIR-центровані мікросервіси як транзакційний каркас інтеграції, копілоти в ЕМК для «ambient»-документації, хмарні LLM

із пошуком та отриманням знань із локальних джерел, а також доменні чат-боти для вузьких клінічних маршрутів. Їх поєднання дає практичний ефект за умови зрілого NLP-конвеєра «вилучення – нормалізація – прив'язка», використання UMLS/SNOMED CT/LOINC/RxNorm, профілювання FHIR та наявності аудиту і клінічного нагляду. Показано, що така композиція знижує навантаження на лікаря, покращує повноту і уніфікацію даних, прискорює підготовку клінічних документів і підвищує перевірюваність відповідей завдяки посиланням на джерела. Для локалізації в Україні критично забезпечити двомовні інтерфейси, зіставлення локальних довідників до міжнародних термінологій, національні профілі FHIR та відповідність вимогам безпеки і юрисдикції обробки даних.

Перспективні напрямки подальших досліджень включають проспективні випробування у первинній ланці, розробку україномовних корпусів для клінічного NLP, стандартизовані бенчмарки для RAG-систем, методики безпечного розмежування доступу та аналізу вартості/ефективності в умовах реального впровадження.

Список літератури:

1. Laranjo L. etc. Conversational agents in healthcare: a systematic review. *Journal of the american medical informatics association.* 2018. Т. 25, № 9. С. 1248–1258. URL: https://doi.org/10.1093/jamia/ocy072 (дата звернення: 04.09.2025).

2. Milne-Ives M. etc. The effectiveness of artificial intelligence conversational agents in health care: systematic review. *Journal of medical internet research*. 2020. Т. 22, № 10. С. e20346. URL: https://doi.org/10.2196/20346 (дата звернення: 04.09.2025).

3. Fan X. etc. Utilization of self-diagnosis health chatbots in real-world settings: case study. *Journal of medical internet research*. 2021. Т. 23, № 1. С. e19928. URL: https://doi.org/10.2196/19928 (дата звернення: 04.09.2025).

4. Abbasian M. etc. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *Npj digital medicine*. 2024. Т. 7, № 1. URL: https://doi.org/10.1038/s41746-024-01074-z (дата звернення: 04.09.2025).

5. Li H. etc. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *Npj digital medicine*. 2023. Т. 6, № 1. URL: https://doi.org/10.1038/s41746-023-00979-5 (дата звернення: 04.09.2025).

6. Zhong W., Luo J., Zhang H. The therapeutic effectiveness of artificial intelligence-based chatbots in alleviation of depressive and anxiety symptoms in short-course treatments: a systematic review and meta-analysis. *Journal of affective disorders.* 2024. URL: https://doi.org/10.1016/j.jad.2024.04.057 (дата звернення: 04.09.2025).

7. Tabari P. etc. State-of-the-Art fhir-based data model and structure implementations: a systematic scoping review (preprint). *JMIR medical informatics.* 2024. URL: https://doi.org/10.2196/58445 (дата звернення: 04.09.2025).

8. Home – subscriptions R5 backport v1.2.0-ballot. Index – FHIR v6.0.0-ballot3. URL: https://build.fhir.org/ig/HL7/fhir-subscription-backport-ig/ (дата звернення: 04.09.2025).

9. Bettoni G. N. etc. Application of HL7 FHIR in a microservice architecture for patient navigation on registration and appointments. 2021 IEEE/ACM 3rd international workshop on software engineering for healthcare (SEH), м. Madrid, Spain, 3 черв. 2021 р. 2021. URL: https://doi.org/10.1109/seh52539.2021.00015 (дата звернення: 04.09.2025).

10. Lukac P. J. etc. A randomized-clinical trial of two ambient artificial intelligence scribe technologies. *MedRxiv.* 2025. URL: https://doi.org/10.1101/2025.07.10.24310414 (дата звернення: 04.09.2025).

11. Tierney A. A. etc. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM catalyst*. 2024. Т. 5, № 3. URL: https://doi.org/10.1056/cat.23.0404 (дата звернення: 04.09.2025).

12. Lawrence K. etc. Informed consent for ambient documentation using generative AI in ambulatory care. *JAMA network open*. 2025. Т. 8, № 7. С. e2522400. URL: https://doi.org/10.1001/jamanetworkopen.2025.22400 (дата звернення: 04.09.2025).

13. Cohen I. G., Ritzman J., Cahill R. F. Ambient listening–legal and ethical issues. *JAMA network open*. 2025. Т. 8, № 2. С. e2460642. URL: https://doi.org/10.1001/jamanetworkopen.2024.60642 (дата звернення: 04.09.2025).

14. Kresevic S. etc. A retrieval augmented generation-based framework for clinical decision making: interpretation of the novel Hepatitis C virus guidelines. *NPJ digital medicine.* 2024. Т. 7. С. 102. URL: https://doi.org/10.1038/s41746-024-01151-8 (дата звернення: 04.09.2025).

15. Yang R. etc. Retrieval-augmented generation for generative artificial intelligence in health care. *Npj health systems*. 2025. Т. 2, № 1. URL: https://doi.org/10.1038/s44401-024-00004-1 (дата звернення: 04.09.2025).

16. Zakka C. etc. Almanac – retrieval-augmented language models for clinical medicine. *Nejm ai*. 2024. Т. 1, № 2. URL: https://doi.org/10.1056/aioa2300068 (дата звернення: 04.09.2025).

17. Shi X. etc. Medical dialogue system: a survey of categories, methods, evaluation and challenges. Findings of the association for computational linguistics ACL 2024, Bangkok, Thailand and virtual meeting. Stroudsburg, PA, USA, 2024. 2840–2861. URL: https://doi.org/10.18653/v1/2024.findings-acl.167 (дата звернення: 04.09.2025).

18. Kraljevic Z. etc. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. *Artificial intelligence in medicine.* 2021. T. 117. C. 102083. URL: https://doi.org/10.1016/j.artmed.2021.102083 (дата звернення: 04.09.2025).

19. Mandel J. C. etc. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *Journal of the american medical informatics association*. 2016. T. 23, № 5. C. 899–908. URL: https://doi.org/10.1093/jamia/ocv189 (дата звернення: 04.09.2025).

20. AuditEvent – FHIR v5.0.0. URL: https://hl7.org/fhir/auditevent.html (дата звернення: 04.09.2025).

21. Lee J. etc. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics.* 2019. URL: https://doi.org/10.1093/bioinformatics/btz682 (дата звернення: 04.09.2025).

22. Singhal K. etc. Toward expert-level medical question answering with large language models. *Nature medicine*. 2025. URL: https://doi.org/10.1038/s41591-024-03423-7 (дата звернення: 04.09.2025).

23. Tu T. etc. Towards conversational diagnostic artificial intelligence. *Nature*. 2025. URL: https://doi.org/10.1038/s41586-025-08866-7 (дата звернення: 04.09.2025).

24. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*. 2004. T. 32, № 90001. C. 267D–270. URL: https://doi.org/10.1093/nar/gkh061 (дата звернення: 04.09.2025).

25. Donnelly K. SNOMED-CT: the advanced terminology for global healthcare. *Studies in health technology and informatics*. 2006. C. 279–290.

26. Vreeman D. J., McDonald C. J., Huff S. M. LOINC: a universal standard for identifying laboratory observations. *Journal of the american medical informatics association*. 2016. T. 23, № 4. C. 641–650.

27. Nelson S. J. etc. Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the american medical informatics association*. 2011. T. 18, № 4. 441–448. URL: https://doi.org/10.1136/amiajnl-2011-000116 (дата звернення: 04.09.2025).

28. Nosratabadi S., Atobishi T., Hegedűs S. Social sustainability of digital transformation: empirical evidence from EU-27 countries. *Administrative sciences*. 2023. T. 13, № 5. C. 126. URL: https://doi.org/10.3390/admsci13050126 (дата звернення: 04.09.2025).

29. Beydokhti R. etc. Ontologies in clinical decision support systems: systematic review. *JMIR medical informatics*. 2023. T. 11. C. e46105.

Kandel K.V. USE OF VIRTUAL ASSISTANTS IN MEDICINE

The work synthesizes approaches to building and integrating virtual assistants into medical information systems based on four complementary architectures. It considers FHIR-centered microservices as a transactional framework for data exchange, Copilot solutions embedded into the electronic medical record for "ambient" documentation during the visit, cloud large language models with retrieval-augmented generation (RAG) over local sources, as well as domain-specific chatbots for narrow clinical scenarios (triage, self-service). It is shown that the quality of virtual assistants is determined by a mature NLP pipeline: clinical fact extraction, normalization, and linking to terminologies (UMLS, SNOMED CT, LOINC, RxNorm) with subsequent creation of structured FHIR resources and auditing. Existing empirical evidence on the benefits of conversational agents in mental health is summarized (meta-analyses of RCTs demonstrate clinically meaningful effects), and requirements for clinician oversight and safety are emphasized. Comparative advantages and limitations of the approaches are provided: microservices ensure scalability and interoperability but complicate operations, Copilot reduces documentation burden yet depends on the EMR vendor and speech-recognition quality, RAG architectures increase the verifiability of answers through citations while requiring strict access control and index freshness, domain-specific bots provide predictable behavior in narrow processes but scale at the cost of proliferating scenarios and rules. A combined approach is substantiated in which FHIR microservices ensure data exchange and integrity, the Copilot operates in the clinician's office, the RAG layer supports search and justification, and specialized bots perform end-to-end actions. Localization steps for Ukraine are outlined through bilingual interfaces, a Ukrainian terminology layer, mapping of reference catalogs to SNOMED CT/LOINC, FHIR profiling for NHSU processes, data-residency requirements, and a complete audit log.

Key words: *microservices, semantic layer, data confidentiality, speech recognition, terminology mapping, architecture.*

Дата надходження статті: 04.09.2025
Дата прийняття статті: 17.09.2025
Опубліковано: 16.12.2025