

## УДОСКОНАЛЕННЯ CRM-СИСТЕМИ ІЗ ВИКОРИСТАННЯМ ВЕЛИКОЇ МОВНОЇ МОДЕЛІ GEMMA 3

<sup>1</sup>Вінницький національний технічний університет, Україна

### Abstract

*Робота присвячена удосконаленню CRM-системи із використанням великої мовної моделі Gemma 3, створенню інтелектуального асистента для автоматизації обробки клієнтських запитів у CRM-системах. Запропоновано архітектуру, що поєднує браузерне розширення із сервером генерації тексту, до якого інтегровано модель. Браузерне розширення забезпечує взаємодію з інтерфейсом CRM, захоплює запити клієнтів у реальному часі та передає їх до сервера, де модель генерує відповідь з урахуванням контексту. Модель налаштована для підтримки природного діалогу, персоналізації відповіді та дотримання корпоративного стилю комунікації. Проведене дослідження показало, що застосування Gemma 3 дозволяє значно знизити навантаження на операторів, підвищити швидкість реагування на запити та забезпечити цілодобову підтримку. Перевагами впровадження є висока гнучкість системи, можливість донавчання моделі на власних даних, а також масштабованість рішення. Отримані результати підтверджують доцільність використання Gemma 3 у CRM-системах для підвищення ефективності обслуговування клієнтів.*

**Ключові слова:** оброблення природної мови, Gemma 3, CRM-система, генерація тексту, штучний інтелект, автоматизація роботи з клієнтами.

### Abstract

*This work is dedicated to enhancing CRM systems through the use of the large language model Gemma 3 and to developing an intelligent assistant for the automation of customer request processing within CRM platforms. An architecture is proposed that combines a browser extension with a text generation server, into which the model is integrated. The browser extension enables interaction with the CRM interface, captures client requests in real time, and forwards them to the server, where the model generates a response considering the context. The model is configured to support natural dialogue, personalize responses, and adhere to the corporate communication style. The conducted research demonstrated that applying Gemma 3 significantly reduces operator workload, improves response speed, and ensures 24/7 customer support. The advantages of the implementation include the system's high flexibility, the ability to fine-tune the model on proprietary data, and the scalability of the solution. The results confirm the feasibility and benefits of using Gemma 3 in CRM systems to enhance the efficiency of customer service.*

**Keywords:** Natural Language Processing, Gemma 3, CRM system, text generation, artificial intelligence, customer service automation.

### Вступ

У сучасному бізнес-середовищі ефективна взаємодія з клієнтами є ключовим фактором успіху компаній. Системи управління взаємовідносинами з клієнтами (CRM) відіграють важливу роль у цьому процесі, забезпечуючи зберігання та оброблення інформації про клієнтів, їхні запити та історію взаємодії [1]. Однак, зі зростанням обсягів даних та кількості клієнтських запитів, виникає потреба в автоматизації обробки інформації для підвищення ефективності роботи операторів та покращення якості обслуговування.

Одним із перспективних напрямків вирішення цієї задачі є впровадження штучного інтелекту (ШІ) у CRM-системи, зокрема використання сучасних мовних моделей для автоматизації оброблення природномовних текстів. Модель Gemma 3, розроблена компанією Google, є новітньою мультимодальною мовною моделлю, яка підтримує понад 140 мов і здатна обробляти як текстові, так і візуальні дані. Це робить її особливо привабливою для застосування в задачах автоматизації відповідей на клієнтські запити [2]. Gemma — це сімейство сучасних мовних моделей від Google з відкритим кодом, створених на базі тих самих напрацювань, що й модель Gemini [3].

Платформа Hugging Face надає інструменти для роботи з мовними моделями, включаючи бібліотеку Transformers, яка спрощує процес інтеграції моделей у різні застосунки. Це відкриває можливості для

створення спеціалізованих ШІ-асистентів, здатних обробляти клієнтські запити в реальному часі, генерувати відповідні відповіді та інтегруватися в існуючі CRM-системи через браузерні розширення або інші інтерфейси.

Таким чином, поєднання можливостей моделі Gemma 3 та інструментів платформи Hugging Face створює передумови для розробки ефективних ШІ-асистентів, які можуть значно покращити процеси оброблення клієнтських запитів, зменшити навантаження на операторів та підвищити загальну ефективність роботи компаній [4].

Метою даного дослідження є удосконалення CRM-системи із використанням великої мовної моделі Gemma 3 та створення інтелектуального асистента для автоматизації оброблення клієнтських запитів у CRM-системах.

### **Ідея розв'язання задачі**

Одним із практичних способів впровадження Gemma 3 в існуючий CRM інтерфейс є використання браузерного розширення. Браузерне розширення може діяти як «міст» між веб-інтерфейсом CRM і мовною моделлю, не вимагаючи при цьому суттєвих змін у бекенді самої CRM-системи. Суть підходу полягає в тому, що розширення вбудовує додатковий скрипт на сторінки CRM, який захоплює контент сторінки та введені користувачем запитання, відправляє їх до мовної моделі та відображає згенеровану відповідь безпосередньо в інтерфейсі [5].

Архітектурно рішення складається з двох частин: фронтенд-складова розширення та бекенд-сервіс для моделі. Фронтенд-розширення вбудовує на сторінці CRM панель або кнопку, що дозволяє користувачу (оператору) поставити запитання або попросити згенерувати відповідь. У разі активації, скрипт може зчитати текст поточного звернення клієнта, виділити ключові деталі (або навіть всю переписку) та сформулювати запит до моделі Gemma 3. Далі цей запит надсилається до бекенду розширення – локального або хмарного сервісу, який виконує виклик моделі.

Отже, браузерне розширення виступає зручним засобом інтеграції Gemma 3 у CRM: воно легко підключається до веб-інтерфейсу, може бути розгорнуте вибірково для окремих користувачів/відділів і не потребує втручання у код самого CRM. Через розширення можливо швидко прототипувати рішення – наприклад, додати бічну панель помічника, що пояснює або підсумовує контекст звернення і пропонує варіанти відповіді оператору.

### **Переваги та недоліки запропонованого підходу**

Використання мовної моделі Gemma 3 через платформу Hugging Face у CRM-системі має низку суттєвих переваг:

- Автоматизація рутинних запитів і підвищення продуктивності. Модель здатна самостійно відповідати на найбільш типові й прості питання клієнтів (статус замовлення, відновлення паролю, інформація про товар тощо) без участі оператора.
- Швидкість реакції та доступність 24/7. Gemma генерує відповіді за лічені секунди або навіть швидше, що різко скорочує час очікування клієнтом. LLM у підтримці забезпечує миттєві відповіді і може працювати цілодобово без перерв, тому клієнти отримують допомогу у будь-який момент, коли звернуться [6]. Це веде до покращення досвіду клієнтів: швидка і завжди доступна підтримка підвищує їхню задоволеність і лояльність до сервісу.
- Послідовність і висока якість відповідей. На відміну від людського фактору, модель завжди дотримується заданого стилю спілкування і стандартів якості.
- Персоналізація та краща залученість. Модель може використовувати наявні дані про клієнта для персоналізації відповідей. Завдяки глибокому аналізу контексту і поведінки клієнта, LLM-боти можуть надавати індивідуалізований досвід обслуговування, що підвищує задоволеність і лояльність.
- Використання накопичених знань та аналітики. Інтегрований у CRM LLM може стати інструментом для накопичення й використання бази знань.
- Відкритість і контроль за рішенням. Оскільки Gemma – відкритою моделлю, організація отримує повний контроль над мовною моделлю та даними.

Попри значні переваги, інтеграція Gemma 3 через Hugging Face у CRM не позбавлена викликів і потенційних недоліків:

- Ризик галюцинацій та неточних відповідей. Як і інші великі мовні моделі, Gemma може іноді генерувати впевнені на вигляд, але фактично неправильні або нерелевантні відповіді – явище,

відоме як «галюцинація» LLM. Це становить небезпеку в контексті обслуговування клієнтів, адже помилкова чи безглузда відповідь може ввести клієнта в оману, зашкодити репутації компанії і підірвати довіру до AI-системи. Дослідження показують, що навіть провідні моделі можуть давати 15–20% відповідей з неточностями [7].

- Питання безпеки даних і конфіденційності. Інтегруючи LLM у CRM, компанія має забезпечити захист чутливої інформації клієнтів. Якщо модель розгорнута через хмарний API, виникає ризик передачі персональних даних назовні.
- Вимоги до обчислювальних ресурсів. Хоча Gemma відносно компактна у порівнянні з гігантськими моделями (GPT-3, GPT-4), її інтеграція все ж потребує суттєвих ресурсів. Інференс LLM – важке завдання для CPU і навіть GPU, особливо при великому контекстному вікні 8k токенів. Впровадження таких моделей у CRM стикається з високим навантаженням на обчислювальні ресурси і потребою у потужному апаратному забезпеченні [10].
- Складність інтеграції та підтримки. Розробка власного рішення (розширення + модель) потребує відповідної експертизи. Потрібно налаштувати середовище Hugging Face, отримати й оновлювати ваги моделі Gemma 3, забезпечити сумісність із браузерами, обробку можливих помилок. Інтерфейс CRM може оновлюватися, через що розширення потрібно буде підтримувати актуальним. Все це – додаткові витрати часу і ресурсів на розробку та тестування, особливо порівняно з підключенням готового SaaS-рішення.
- Обмеження домену та необхідність навчання. Базова модель Gemma 3 навчена на широкому корпусі даних, але може не знати специфічних термінів або продуктів компанії. Без додаткового налаштування вона може давати узагальнені відповіді. Тому практично завжди потрібно проводити донавчання (fine-tuning) або реалізовувати підвантаження корпоративних знань.
- Відповідність етичним нормам і контроль виходу. Модель повинна поводитися коректно: не видавати образливих або невідповідних висловлювань, не розголошувати заборонену інформацію. Хоча Gemma 3 має певні механізми безпеки, відкритість моделі означає, що відповідальність за фільтрацію контенту лежить на інтеграторові.

### Вибір архітектури рішення

Помічник підтримки на основі Gemma 3 у CRM через розширення. Розглянемо узагальнену архітектуру, яка впливає з попереднього опису (рис. 1). На комп'ютерах співробітників встановлено браузерне розширення «CRM Assistant». Коли оператор відкриває картку звернення клієнта в CRM (веб-сторінка), розширення вбудовує у її інтерфейс кнопку «Згенерувати відповідь» та бічну панель для взаємодії з Gemma. Текст звернення автоматично витягується із сторінки і підставляється у прихований контекст. При натисканні кнопки, розширення формує запит: наприклад, промпт виду «Клієнт запитує: <текст запиту>. Сформулюй ввічливу професійну відповідь від імені агента підтримки.». Цей промпт надсилається через бекенд (розгорнутий на внутрішньому сервері компанії) до моделі Gemma, яка була завантажена із Hugging Face (або розгорнута як контейнер). Модель генерує відповідь і надсилає її назад розширенню, яке відображає текст у бічній панелі. Оператор перевіряє та, за потреби, редагує її та вставляє у CRM для відправки клієнту. Такий цикл економить час оператора на написання відповіді з нуля. Додатково, панель може мати опції: «Уточнити», «Скоротити відповідь», «Змінити тон» – що виконуються також через додаткові промпти до Gemma (наприклад, «Скороти наведену вище відповідь до 1 абзацу»). Цей приклад показує, як LLM інтегрується у робочий процес підтримки безпосередньо на інтерфейсі, використовуючи Hugging Face як платформу для моделі і браузерне розширення як засіб інтеграції.

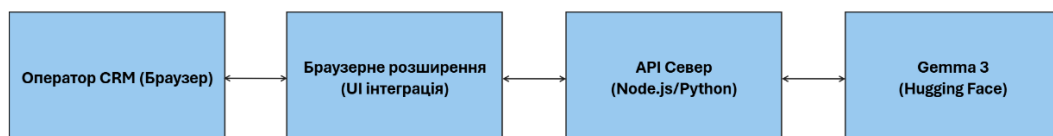


Рисунок 1. Діаграма розгортання Gemma 3 у CRM-системі через браузерне розширення

Платформа Hugging Face відіграє ключову роль у прискоренні впровадження подібних рішень. Вона надає готову інфраструктуру для зберігання і розповсюдження моделей, інструменти для

налаштування та оптимізації, а також спільноту, де можна знайти приклади успішної інтеграції. Усе це дозволяє провести шлях від ідеї до прототипу інтеграції LLM у CRM значно швидше. Як результат, компанії можуть експериментувати з відкритими мовними моделями, такими як Gemma, отримуючи вигоди від AI в обслуговуванні клієнтів і, водночас, зберігаючи контроль над своїми даними та процесами.

### Висновки

У роботі запропоновано як мовна модель Gemma може бути впроваджена в систему управління відносинами з клієнтами (CRM) для автоматизованого оброблення звернень. Розглянути сценарії застосування Gemma 3 для оброблення клієнтських запитів у режимі реального часу та генерування відповідей безпосередньо в інтерфейсі CRM, з урахуванням авторського досвіду. Проаналізовано переваги та недоліки такого підходу, а також вимоги, продуктивність та безпекові аспекти рішення.

Впровадження мовної моделі Gemma 3 у CRM – технічно здійсненний та багатообіцяючий підхід. Він забезпечує автоматизацію та покращення якості підтримки клієнтів, пропонуючи рішення в реальному часі і масштабуючись під зростаючі потреби бізнесу. Відкрита природа Gemma дає гнучкість у налаштуванні та розгортанні, а правильно спроектована інтеграція (наприклад, через браузерне розширення з урахуванням всіх безпекових вимог) дозволяє безпечно і ефективно використати її можливості. Це підтверджує тренд, що великі мовні моделі стають невід’ємною частиною сучасних CRM-систем, підвищуючи їх інтелектуальність і цінність для бізнесу.

### СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Як CRM система систематизує спілкування з клієнтами в месенджерах [Електронний ресурс] – Режим доступу: <https://creativesmm.com.ua/iak-crm-systema-systematyzuie-spilkuvannia-z-kliientamy-v-mesendzherakh/>
2. Introducing Gemma 3: The most capable model you can run on a single GPU or TPU [Електронний ресурс] – Режим доступу: <https://blog.google/technology/developers/gemma-3/>
3. Access Gemma on Hugging Face [Електронний ресурс] – Режим доступу: <https://huggingface.co/google/gemma-7b>
4. Welcome Gemma 3: Google's all new multimodal, multilingual, long context open LLM [Електронний ресурс] – Режим доступу: <https://huggingface.co/blog/gemma3>
5. kongchenglc/LlamaChromeSidebar [Електронний ресурс] – Режим доступу: <https://github.com/kongchenglc/LlamaChromeSidebar>
6. How to find the best Open-Source LLM for your Customer Service Chatbot [Електронний ресурс] – Режим доступу: <https://www.giskard.ai/knowledge/find-the-best-llm-for-your-customer-service-chatbot>
7. Hallucinations in LLMs: What You Need to Know Before Integration [Електронний ресурс] – Режим доступу: <https://masterofcode.com/blog/hallucinations-in-llms-what-you-need-to-know-before-integration>

**Пуженко Віталій Валерійович** – студент групи ICT-24м, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м.Вінниця, e-mail: [vitalii.puzhenko@gmail.com](mailto:vitalii.puzhenko@gmail.com)

**Мокін Віталій Борисович** – д-р техн. наук, професор, завідувач кафедри системного аналізу та інформаційних технологій; e-mail: [ybmokin@vntu.edu.ua](mailto:ybmokin@vntu.edu.ua)

**Puzhenko Vitalii V.** – student of ICT-24m, Faculty of Intellectual Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: [vitalii.puzhenko@gmail.com](mailto:vitalii.puzhenko@gmail.com)

**Mokin Vitalii B.** – Dr. Sc. (Eng.), Professor, Head of the Chair of System Analysis and Information Technologies, e-mail: [ybmokin@vntu.edu.ua](mailto:ybmokin@vntu.edu.ua)