УДК 004.773.6:004.032.26]:334.72
DOI: 10.31673/2409-7292.2024.040014

Кушнерьов О. С., Вадько К. С., Кузченко А. О.

ЧАТ-БОТ НА ОСНОВІ НЕЙРОМЕРЕЖІ ДЛЯ НАДАННЯ ІНФОРМАЦІЙНОЇ ПІДТРИМКИ БІЗНЕСУ З ПОЗИЦІЇ КІБЕРБЕЗПЕКИ

У статті розглянуто процес розробки інтелектуального чат-бота для бізнесу, що надає інформаційну підтримку з урахуванням вимог кібербезпеки. Акцент зроблено на ролі чат-ботів, побудованих на основі нейронних мереж, у сучасному бізнес-середовищі. Зокрема, досліджено їхню здатність автоматизувати обробку клієнтських запитів, оптимізувати взаємодію з користувачами та підвищувати якість обслуговування клієнтів. Детально проаналізовано ключові кіберзагрози, пов'язані з використанням чат-ботів, зокрема ті, що описані в OWASP Top 10 для великих мовних моделей (LLM). Серед основних вразливостей розглянуто ризики витоку даних, зловживання функціональністю та маніпулювання запитами. Вивчено методи захисту, що включають автентифікацію та авторизацію користувачів, наскрізне шифрування даних, застосування самознищуваних повідомлень, а також забезпечення відповідності вимогам GDPR для захисту персональних даних. Окрему увагу приділено інтеграції механізмів кіберзахисту в архітектуру чат-ботів. Підкреслено важливість поєднання технологій штучного інтелекту та сучасних інструментів кібербезпеки для підвищення стійкості бізнес-процесів до загроз. У висновках відзначено, що створення безпечних чат-ботів сприяє ефективності роботи компаній та забезпеченню високого рівня захисту інформації в умовах цифрової трансформації. Така інтеграція дозволяє мінімізувати ризики та розширює можливості застосування чат-ботів у різних галузях, що робить їх важливим інструментом для забезпечення конкурентоспроможності в умовах швидких технологічних змін.

Ключові слова: чат-боти, нейронні мережі, кібербезпека, інформаційна підтримка, автоматизація бізнес-процесів, кіберзагрози, великі мовні моделі (LLM).

Вступ

У сучасному бізнес-середовищі інформаційна підтримка виступає одним із основних чинників успішного функціонування підприємств. Компанії все частіше стикаються з необхідністю оптимізувати комунікаційні процеси, що не лише сприяє покращенню взаємодії з клієнтами, але й сприяє вдосконаленню внутрішнього обміну інформацією, що, у свою чергу, підвищує надійність і безпеку бізнес-процесів. У цьому контексті інтелектуальні технологічні рішення, такі як чат-боти, які функціонують на основі нейронних мереж, вважаються перспективними інструментами для автоматизації обробки клієнтських запитів. Вони дозволяють оперативно вирішувати питання та суттєво покращують якість обслуговування клієнтів.

Нейронні мережі забезпечують широкі можливості для розробки адаптивних чат-ботів, особливо у таких сферах, як обробка природної мови (Natural Language Processing, NLP) та машинне навчання. Завдяки глибинному навчанню ці боти здатні самостійно навчатися на великих обсягах даних, що дозволяє їм адаптувати відповіді відповідно до специфічних потреб користувачів. Високий рівень адаптивності таких рішень сприяє персоналізації обслуговування та підвищує ефективність взаємодії з клієнтами, водночас зменшуючи навантаження на працівників компанії. Водночас застосування таких технологій породжує значні виклики у сфері кібербезпеки.

Використання чат-ботів на базі нейронних мереж, особливо у бізнес-середовищі, супроводжується низкою кібербезпекових викликів. Основні ризики включають крадіжку персональних даних, атаки на канали зв'язку, а також можливість підробки запитів. Вразливості у таких системах можуть призвести до серйозних втрат як для компанії, так і для її клієнтів, а також до зниження рівня довіри з боку користувачів. Відтак, для забезпечення безпеки інформаційної підтримки чат-ботами необхідно запроваджувати ефективні кіберзахисні заходи.

Метою цієї статті є дослідження процесу розробки інтелектуального чат-бота для бізнесу, який забезпечує інформаційну підтримку з дотриманням вимог кібербезпеки. На основі аналізу вразливостей, що загрожують інформаційній безпеці чат-ботів, у статті розглянуто ключові аспекти забезпечення захисту, зокрема шифрування даних, автентифікацію користувачів та використання систем виявлення кіберзагроз. Такі методи є

критично важливими для забезпечення захисту інформації від несанкціонованого доступу та гарантування надійності обміну даними.

Дослідження показує, що інтеграція нейронних мереж у чат-боти разом із впровадженням кіберзахисних механізмів дозволяє не лише підвищити рівень автоматизації бізнес-процесів, а й забезпечити високий рівень захисту від потенційних кіберзагроз. Така інтеграція набуває особливого значення в умовах цифрової трансформації, яка охоплює більшість галузей економіки, підвищуючи актуальність питань інформаційної безпеки.

Аналіз літературних джерел та формулювання проблеми

Зі зростанням потреби в автоматизації бізнес-процесів чат-боти, побудовані на основі нейронних мереж, стали невід'ємною частиною сучасного бізнес-середовища. Ряд досліджень наголошують на перевагах застосування чат-ботів у різних галузях, включаючи покращення обслуговування клієнтів, оперативність відповідей та зменшення навантаження на персонал [1, 2]. Однак, у контексті зростання використання великих мовних моделей у цих системах виникають нові виклики в сфері кібербезпеки. Чат-боти здатні не тільки імітувати людське спілкування, але й обробляти значний обсяг даних, що підвищує ризик витоку або маніпуляції цими даними [3, 4].

Дослідження показують, що однією з ключових проблем, пов'язаних з безпекою чат-ботів, є вразливості великих мовних моделей. Зокрема, OWASP визначає низку загроз для таких систем, включаючи атаки на введення даних (Prompt Injection) та маніпуляції навчальними даними (Training Data Poisoning), які можуть призводити до небажаної поведінки ботів або навіть витоку конфіденційної інформації [5, 10–12]. Проблема ускладнюється тим, що чат-боти часто взаємодіють із зовнішніми користувачами, що робить їх особливо вразливими до неконтрольованих загроз. Неврахування цих ризиків може негативно вплинути на довіру користувачів до чат-ботів і загалом на ефективність їхнього застосування в бізнесі [6].

Інша суттєва загроза пов'язана з недостатнім дотриманням вимог захисту даних, таких як GDPR, особливо коли чат-боти обробляють особисту інформацію користувачів. Дослідники акцентують на важливості використання багатофакторної автентифікації, наскрізного шифрування та обмеження часу зберігання даних для мінімізації ризиків [7]. Відомо, що чат-боти, розроблені на основі нейронних мереж, мають високу адаптивність, що дозволяє їм автоматично навчатися на основі зібраних даних. Однак, відсутність належних механізмів контролю за доступом до цих даних може спричинити значні втрати як для компанії, так і для її клієнтів [8].

Таким чином, головною проблемою є баланс між функціональністю та безпекою чат-ботів. Використання нейронних мереж підвищує їхню адаптивність і здатність до самонавчання, що є важливим для забезпечення персоналізованої комунікації. Однак, недостатня увага до безпеки чат-ботів може призвести до серйозних загроз, включаючи атаки на особисті дані користувачів, несанкціонований доступ і витоки конфіденційної інформації. Це вимагає інтеграції кіберзахисних механізмів на всіх етапах створення та експлуатації чат-ботів, забезпечуючи при цьому відповідність нормативним вимогам та захист особистих даних користувачів [3, 9, 13–15].

Проблема забезпечення кібербезпеки чат-ботів є складною, адже вони взаємодіють як з внутрішніми системами компаній, так і з зовнішніми користувачами. Розробка ефективних стратегій захисту, що враховуватимуть усі потенційні вразливості та дотримуватимуться вимог законодавства, є критично важливим кроком для надійного використання чат-ботів у бізнес-середовищі.

Мета роботи та цілі дослідження

Метою даного дослідження є розробка та аналіз інтелектуального чат-бота для бізнес-середовища, який забезпечує інформаційну підтримку з високим рівнем кібербезпеки. Основна увага приділяється дослідженню можливостей нейронних мереж для підвищення адаптивності та функціональності чат-бота, а також вивченню методів захисту інформації,

© Кушнерьов, О. С., Вадько, К. С., & Кузченко, А. О. (2024). Чат-бот на основі нейромережі для надання інформаційної підтримки бізнесу з позиції кібербезпеки. Сучасний захист інформації, 4(60), 131–140. https://doi.org/10.31673/2409-7292.2024.040014.

зокрема шифрування даних, автентифікації користувачів та використання систем виявлення кіберзагроз.

Для вирішення поставленої мети розглянуто такі завдання:

1. Аналіз вразливостей чат-ботів, які функціонують на основі нейронних мереж, та ризиків для інформаційної безпеки.
2. Розгляд методів захисту, що застосовуються для забезпечення безпеки обміну даними між користувачами та чат-ботами.
3. Розробка рекомендацій для підвищення кібербезпеки чат-ботів, які працюють у бізнес-середовищі, з урахуванням вимог GDPR та інших регуляторних актів.
4. Впровадження надійних засобів кібербезпеки для мінімізації ризиків, пов'язаних з автоматизованими системами обробки клієнтських запитів.

Інтелектуальні чат-боти для бізнесу в аспекті функціональних можливостей та ризиків кіберзагроз

Чат-бот — це сучасний програмний інструмент, здатний імітувати людське спілкування шляхом обробки текстових повідомлень та надання відповідей на запити користувачів. Інтерактивний та адаптивний за своєю суттю, чат-бот може не лише відповідати на питання, але й виконувати різноманітні дії на запит користувача, а також забезпечувати розважальний контент. Основу таких систем становить обробка природної мови (NLP), яка дозволяє чат-ботам розуміти вхідні повідомлення, аналізувати їхній зміст і генерувати відповідні відповіді. Чат-боти можуть функціонувати як автономні програми або інтегруватися у функціонал інших систем, наприклад, пошукових платформ чи соціальних мереж [3].

Сучасні чат-боти використовують штучний інтелект (AI) для більш ефективної та гнучкої взаємодії з користувачами. Інтеграція AI дозволяє автоматизувати інтелектуальну поведінку, надаючи машинам здатність імітувати людиноподібну комунікацію. Застосування технологій обробки природної мови (NLP), мови розмітки штучного інтелекту (AIML), зіставлення шаблонів та технологій розуміння природної мови (NLU) дозволяє чат-ботам відповідати на запити користувачів, інтерпретувати їхні потреби та забезпечувати функції, що варіюються від інформаційної підтримки до маркетингових активностей. Чат-боти також забезпечують можливість персоналізованої комунікації, підтримуючи таргетовану взаємодію, що робить їх ефективним інструментом для взаємодії з конкретними аудиторіями [2].

Чат-боти можна класифікувати за кількома критеріями, зокрема за областю знань, типом послуг, поставленими цілями, методами обробки даних, генерації відповідей, ступенем залучення людини та методами побудови. Основні напрями їхнього застосування включають електронну комерцію, обслуговування клієнтів, виконання функцій віртуального помічника, протидію шахрайству та моніторинг об'єктів. У сфері електронної комерції чат-боти забезпечують додатковий канал комунікації з клієнтами, сприяючи швидшій обробці замовлень і запитів, що підвищує рівень задоволеності споживачів. У сфері обслуговування клієнтів чат-боти зазвичай виступають першою лінією контакту, забезпечуючи швидке вирішення запитів і знижуючи навантаження на працівників компанії. Як віртуальні помічники, вони допомагають користувачам у вирішенні повсякденних завдань, таких як управління розумним будинком або планування подій.

Чат-боти також застосовуються для протидії шахрайству та моніторингу підозрілих активностей, надаючи компаніям додатковий рівень захисту від кібератак і сприяючи захисту конфіденційної інформації. Крім того, деякі боти виконують моніторинг у режимі реального часу, що дозволяє швидко реагувати на потенційні загрози, навіть коли користувач неактивний [4].

Розвиток технологій машинного навчання та обробки природної мови значно сприяв удосконаленню чат-ботів, завдяки чому з'явилися нові типи, здатні краще адаптуватися до потреб користувачів. Це, своєю чергою, сприяло зростанню популярності чат-ботів серед компаній, які шукають шляхи підвищення ефективності діяльності шляхом автоматизації процесів обслуговування клієнтів [9].

Проте зростання використання чат-ботів також супроводжується збільшенням кількості потенційних загроз та вразливостей у сфері кібербезпеки. Основні кіберзагрози включають такі атаки, як підробка даних, відмова в обслуговуванні та несанкціонований доступ до конфіденційної інформації. Для зниження рівня ризиків важливо впроваджувати надійні заходи безпеки, зокрема наскрізне шифрування комунікацій та багатофакторну автентифікацію користувачів. Це дозволяє захистити дані від потенційних загроз, зберігаючи при цьому функціональність і зручність використання чат-ботів у бізнес-середовищі.

Загрози та вразливості є основними факторами, що впливають на кібербезпеку чат-ботів. Кіберзагрози можна визначити як методи, за допомогою яких зловмисники можуть здійснити злам комп'ютерної системи. Прикладами таких загроз для чат-ботів є підробка даних, відмова в обслуговуванні, несанкціоноване розкриття інформації та підвищення привілеїв. Вразливості, навпаки, є слабкими місцями системи, які можуть бути використані для компрометації, якщо їх не захищено належним чином. Недостатнє обслуговування системи, слабке кодування, відсутність належного захисту та людські помилки підвищують уразливість системи до атак. Для мінімізації вразливостей можна застосовувати самознищувані повідомлення у поєднанні з іншими заходами безпеки, такими як наскрізне шифрування, захищений протокол, автентифікація та авторизація користувачів. Додатковим засобом забезпечення безпеки чат-ботів є використання аналітики поведінки користувачів (UBA) [2].

Чат-боти можна умовно поділити на дві основні групи: скриптові та розумні. Скриптові чат-боти працюють на основі попередньо визначених команд і сценаріїв, що обмежує їхні можливості. Під час взаємодії з такими ботами користувач обирає одну з заздалегідь запрограмованих опцій, через що реакція скриптових чат-ботів залишається обмеженою визначеними сценаріями, що знижує їхню адаптивність [6]. Натомість розумні чат-боти використовують технології штучного інтелекту та обробки природної мови (NLP), які дозволяють їм навчатися на основі отриманих даних, розуміти контекст і адаптуватися до потреб користувача.

Розумні чат-боти можна розділити на два типи: генеративні та пошукові. Генеративні боти використовують методи машинного навчання (ML) для автоматичного створення відповідей. Ці чат-боти працюють з великим обсягом навчальних даних, що дозволяє їм формувати відповіді на запити користувачів без опори на заздалегідь підготовлену базу відповідей [6]. Алгоритм роботи генеративного чат-бота представлено на рис. 1. Завдяки таким можливостям генеративні боти є більш інтелектуальними, оскільки можуть адаптуватися до різноманітних запитань і надавати гнучкі, відповідні до контексту відповіді, що значно підвищує їхню ефективність та корисність для користувачів.

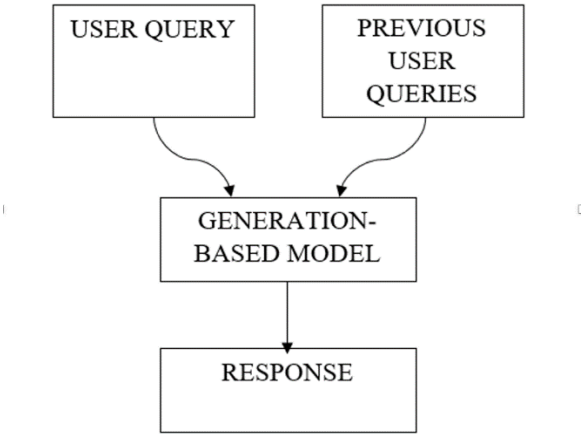

Рис. 1. Алгоритм роботи чат-боту на основі генерації [11]

Пошукові чат-боти, або боти, які працюють на основі пошукових алгоритмів, обробляють отримані повідомлення за допомогою текстового та контекстного аналізу, що дозволяє їм обирати відповіді з існуючої бібліотеки заздалегідь підготовлених відповідей. Контекст повідомлення може включати поточне місцезнаходження користувача в структурі діалогу (дереві діалогу) та всі попередні повідомлення, надіслані в межах конкретної сесії. Це дозволяє ботам не лише відповідати на конкретні запити, а й враховувати попередній контекст, що підвищує актуальність і точність відповідей.

Якщо пошукові чат-боти були попередньо навчені на великому й різноманітному наборі даних, який охоплює численні сценарії запитів та відповідей, їх здатність генерувати релевантні відповіді суттєво підвищується. Навчання на основі обширного корпусу даних сприяє тому, що такі боти рідко припускаються помилок у своїх відповідях. Це зумовлено тим, що обсяг і варіативність навчального набору даних дозволяють пошуковому алгоритму “навчитися” ефективно розпізнавати контексти, які відповідають певним запитам, та обирати оптимальні відповіді навіть для схожих або частково незвичних формулювань.

Проте варто зазначити, що робота пошукових чат-ботів має свої обмеження. Наприклад, якщо користувач ставить запитання, на яке в бібліотеці відповідей немає заздалегідь підготовленої відповіді, бот не здатен надати змістовну відповідь. Подібні обмеження виникають також у випадках, коли бот не може правильно проаналізувати специфічне або надто складне повідомлення користувача через неоднозначність його змісту чи надмірну кількість контекстних факторів, які алгоритм не може врахувати. У таких випадках бот може відобразити загальне повідомлення про помилку або перенаправити користувача до інших доступних джерел інформації [11].

Алгоритм функціонування чат-бота на основі пошуку подано на рис. 2, де показано ключові етапи процесу: від аналізу отриманого повідомлення до вибору відповіді з бібліотеки та формування кінцевого результату.

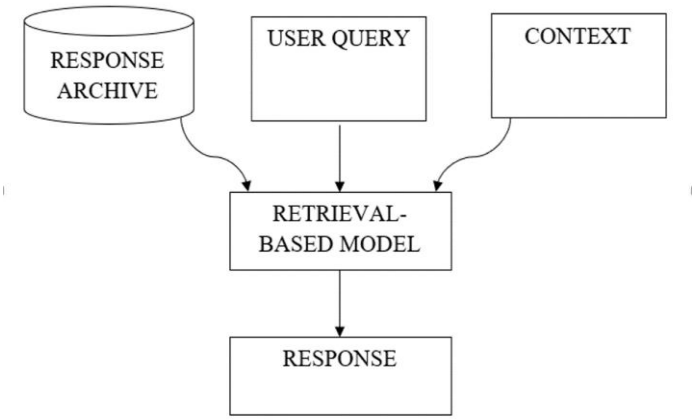

Рис. 2. Алгоритм роботи чат-боту на основі пошуку [11]

Чат-боти різних типів мають свої переваги та недоліки, що робить їх універсальними інструментами для компаній, які прагнуть розвиватися в онлайн-середовищі. Кожен тип чат-бота має унікальні характеристики, які визначають його основні функції та сферу застосування, дозволяючи використовувати їх у різних галузях бізнесу, де комунікація з клієнтами відіграє ключову роль. На сучасному ринку чат-боти відіграють важливу роль у підвищенні ефективності бізнес-процесів, сприяючи автоматизації та оптимізації завдань, які раніше вимагали залучення співробітників, зокрема для консультування клієнтів. Впровадження чат-ботів дозволяє компаніям забезпечувати безперервне обслуговування клієнтів завдяки оперативному та ефективному реагуванню на їхні запити, що сприяє покращенню загального рівня обслуговування. Чат-боти також оптимізують процеси купівлі

та консультування, забезпечуючи клієнтів швидким доступом до необхідної інформації та підтримки, що робить процес купівлі зручнішим. Крім того, автоматизація повторюваних завдань дозволяє підвищити продуктивність, звільняючи людські ресурси від рутинних обов'язків і даючи їм змогу зосередитися на складніших і важливіших завданнях [6].

Сучасні компанії активно впроваджують чат-боти в бізнес-процеси, постійно розширюючи їх функціональні можливості для покращення обслуговування клієнтів. Чат-боти відіграють важливу роль у забезпеченні ефективної комунікації між компанією та її клієнтами, оптимізуючи взаємодію та знижуючи навантаження на контакт-центри. Існують різні типи чат-ботів, кожен з яких має свої унікальні можливості та сфери застосування в бізнес-середовищі. Голосові боти забезпечують інтерактивну комунікацію, поєднуючи текстове та голосове спілкування із застосуванням технологій штучного інтелекту та обробки природної мови (NLU). Завдяки здатності розпізнавати мовні сигнали й конвертувати текст у мовлення (TTS), голосові помічники ефективно інтегруються у різні сервіси, сприяючи покращенню обслуговування клієнтів. Гібридні чат-боти поєднують автоматизацію з можливістю переходу до живого оператора. У випадку, коли автоматизована система не може надати відповідь, запит автоматично передається представнику служби підтримки, що дозволяє надавати більш персоналізоване обслуговування.

Чат-боти для соціальних мереж спеціалізуються на роботі у платформах соціальних мереж та месенджерах, використовуючи алгоритми штучного інтелекту для підвищення ефективності взаємодії з клієнтами. Це сприяє зниженню навантаження на контакт-центри та забезпечує безперервне обслуговування. Чат-боти на основі меню працюють на фіксованій структурі дерева рішень, що дозволяє клієнтам отримувати відповіді через вибір із запропонованих варіантів, проте жорстка структура таких ботів може обмежувати їх ефективність у складних запитах. Чат-боти на основі ключових слів застосовують методи обробки природної мови (NLP) для розуміння запитів за допомогою ключових слів, проте часте використання цього підходу може знижувати точність відповідей. Чат-боти на основі правил працюють за логікою if/then, що забезпечує гнучкість у вирішенні різних запитів через аналіз слів та їх структури.

Контекстні чат-боти з елементами штучного інтелекту здатні зберігати контекст бесіди та використовувати його в подальших взаємодіях. Такі боти часто інтегруються із CRM-системами, що дозволяє їм збирати додаткову інформацію про клієнтів, підвищуючи ефективність персоналізованої взаємодії. Чат-боти підтримки орієнтовані на допомогу клієнтам після основного обслуговування і зазвичай використовуються на платформах онлайн-документації або для самообслуговування [3]. Таким чином, чат-боти стають важливим інструментом для підвищення рівня обслуговування клієнтів у різних сферах бізнесу. Вибір конкретного типу чат-бота залежить від специфіки завдань та потреб клієнтів. Незважаючи на численні переваги, слід враховувати можливі обмеження кожного типу, які можуть впливати на ефективність комунікації. Подальший розвиток технологій штучного інтелекту та обробки природної мови (NLP) сприятиме вдосконаленню функціональності чат-ботів і підвищенню їх корисності для клієнтського обслуговування.

У сучасному світі великі мовні моделі (LLM) стали важливою складовою різноманітних систем, зокрема в електронній комерції, обслуговуванні клієнтів та інших сферах. Однак разом із їх активним впровадженням зростає кількість вразливостей, що створюють потенційні загрози для безпеки користувачів і організацій. Вивчення основних вразливостей, зазначених у OWASP Top 10 для великих мовних моделей, є необхідним кроком для зниження ризиків та забезпечення безпеки сучасних систем, що використовують LLM [10]. OWASP Top 10 підкреслює десять ключових вразливостей, які слід враховувати при розгортанні та експлуатації LLM.

Перша вразливість, Prompt Injection (LLM01), стосується ризику маніпуляцій мовними моделями через спеціально створені введення. Зловмисні ін'єкції можуть спричинити несанкціонований доступ, витоки даних та ухвалення помилкових рішень [10]. Неналежна

обробка введення може призвести до витоку конфіденційної інформації або виконання шкідливого коду, що значно знижує надійність системи.

Друга вразливість, Insecure Output Handling (LLM02), акцентує увагу на необхідності ретельної перевірки результатів, що генеруються мовними моделями. Без належної валідації вихідні дані можуть бути використані для подальших атак, таких як виконання коду, що ставить під загрозу системи та призводить до витоків даних [10]. Це вимагає розробки відповідних механізмів фільтрації та контролю за результатами, що генеруються LLM, особливо в умовах автономної взаємодії з іншими системами.

Третя вразливість, Training Data Poisoning (LLM03), пов'язана із загрозою отруєння навчальних даних. Модифіковані або навмисно викривлені дані можуть призвести до небезпечних або неетичних відповідей LLM, а також знижувати точність і безпечність моделі [10]. Це підкреслює необхідність регулярного моніторингу джерел даних та забезпечення їхньої цілісності й надійності.

Вразливість Model Denial of Service (LLM04) пов'язана з ризиком перевантаження моделей запитами, які вимагають значних ресурсів. Це може призвести до збоїв у наданні послуг і суттєвих витрат на їх відновлення [10]. Для захисту від подібних атак доцільно обмежити доступ до ресурсомістких функцій та встановити ліміти на обробку запитів. Supply Chain Vulnerabilities (LLM05) вказує на загрозу безпеці, пов'язану з можливістю використання скомпрометованих компонентів, сервісів або даних, що можуть знизити цілісність системи [10]. Недостатня увага до безпеки ланцюга постачання може призвести до витоків даних або повного збою системи, тому важливо враховувати ці аспекти при розробці надійних LLM-рішень. Sensitive Information Disclosure (LLM06) підкреслює важливість запобігання випадковому розкриттю конфіденційної інформації у результатах роботи LLM. Незахищеність цих даних може призвести до правових наслідків і втрати конкурентної переваги [10]. Забезпечення захисту приватності має бути пріоритетом у будь-якій системі, що використовує LLM. Наступною вразливістю є Insecure Plugin Design (LLM07), де незахищені плагіни, що обробляють ненадійні введення, можуть створювати ризики, такі як віддалене виконання коду. Для мінімізації ризиків важливо впроваджувати чіткі механізми контролю доступу та обмеження, які допомагають знизити ймовірність експлуатації плагінів [10].

Excessive Agency (LLM08) відображає ризики, пов'язані з надмірною автономією, наданою LLM. Це може призвести до неконтрольованих дій, які ставлять під загрозу надійність, приватність і довіру до системи [10]. Тому доцільно обмежити обсяги доступу моделі до чутливих операцій і контролювати її взаємодію з іншими компонентами.

Вразливість Overreliance (LLM09) стосується ризику надмірного покладання на результати роботи мовної моделі без їх критичної оцінки. Це може призвести до ухвалення помилкових рішень, вразливостей у безпеці та правової відповідальності [10]. Щоб уникнути цього, важливо використовувати результати LLM як допоміжний інструмент, а не як єдине джерело для прийняття рішень.

Нарешті, Model Theft (LLM10) підкреслює важливість захисту мовних моделей від несанкціонованого доступу. Незахищеність може призвести до крадіжки інтелектуальної власності, втрати конкурентних переваг і розповсюдження конфіденційної інформації [10]. Це ставить на порядок денний необхідність впровадження надійних засобів аутентифікації та контролю доступу для захисту LLM.

Забезпечення безпеки великих мовних моделей є комплексним завданням, що вимагає багатостороннього підходу до управління ризиками. Особливо важливим є використання захисних механізмів на всіх етапах життєвого циклу моделі, починаючи з навчання і закінчуючи впровадженням, з метою збереження конфіденційності, цілісності та доступності даних. Глибоке розуміння основних вразливостей, механізмів їхньої експлуатації та можливих наслідків є основою для розробки надійних і ефективних стратегій захисту, що стає особливо актуальним на фоні постійно зростаючих загроз у сфері кібербезпеки. Сучасний розвиток великих мовних моделей, таких як LLM, значною мірою спирається на використання

потужних бібліотек і фреймворків, таких як Hugging Face Transformers, OpenAI GPT, BERT, XLNet, TensorFlow та PyTorch. Ці інструменти стали основними у створенні та навчанні мовних моделей, значно спрощуючи процес розробки та надаючи готові реалізації й засоби для дослідження. Однак разом із перевагами використання цих фреймворків виникає низка вразливостей, зокрема пов'язаних із функціями та параметрами, які можуть бути критичними для безпеки моделей.

У мовах програмування, таких як Python, існують функції, які можуть створити серйозні загрози для безпеки при неналежному використанні. Наприклад, функції `eval()` та `exec()`, що дозволяють виконувати довільний код, значно підвищують ризик ін'єкції шкідливого коду, тому їх використання має бути обмеженим і суворо контрольованим. Подібну загрозу становить модуль pickle, призначений для серіалізації об'єктів, оскільки неконтрольована десеріалізація може призвести до виконання небезпечного коду. Використання функції subprocess, яка може спричинити виконання небезпечних команд, зокрема віддалене виконання коду (RCE), також вимагає ретельної перевірки.

Інші функції, такі як `input()` і `os.system()`, також можуть бути потенційно небезпечними. Функція `input()`, що приймає дані від користувача, потребує належної валідації для уникнення атак, тоді як `os.system()`, виконуючи команди системної оболонки, також створює загрозу RCE у випадку відсутності контролю введення. Доступ до URL через функцію `urllib.urlopen()` може призвести до атак типу Server-Side Request Forgery (SSRF), а неконтрольоване використання `sqlite3.execute()` може викликати SQL-ін'єкцію, що є серйозною загрозою для цілісності бази даних. Динамічний імпорт модулів за допомогою `import()` також створює ризик ін'єкції шкідливого коду, якщо не забезпечується належний контроль, тоді як недостатня перевірка форматування у `format()` може призвести до віддаленого виконання коду та інших проблем із безпекою. Функція `requests.get()` також є вразливою до атак типу Cross-Site Request Forgery (CSRF) за неналежного контролю, що ставить під загрозу цілісність системи [5].

Враховуючи наведені вразливі функції, стає очевидним, що забезпечення безпеки великих мовних моделей вимагає підвищеної уваги. Кожна з перерахованих функцій може слугувати вхідною точкою для атак, що ставлять під загрозу цілісність і конфіденційність даних. Для зменшення ризиків необхідно впроваджувати відповідні практики безпеки, такі як валідація введення, контроль за виконанням кодів та обмеження доступу до небезпечних функцій. Це сприятиме створенню більш безпечних і надійних систем на базі LLM.

Вразливості системи — це слабкі місця, які можуть бути використані зловмисниками для несанкціонованих дій, таких як порушення привілеїв у комп'ютерній системі [1]. Система стає вразливою, якщо має слабке кодування, застаріле апаратне забезпечення, ненадійний брандмауер тощо. Часто вразливості виникають через людські помилки. Security Development Lifecycle (SDL) допомагає уникати таких помилок. Оскільки багато чат-ботів використовують хмарні обчислювальні служби, що мають свої ризики та вразливості, важливо належним чином захищати ці дані. Подальший текст зосереджується на комунікаційній безпеці та аспектах обробки даних. Компанії, що не використовують хмарні служби, часто звертаються до цих питань під час інтеграції систем чат-ботів.

Безпечний обмін повідомленнями складається з двох ключових доменів. Перший домен охоплює безпеку передачі даних (зокрема, повідомлень, голосових записів, зображень) на сервер чат-бота, тоді як другий домен стосується безпеки обробки, зберігання та передачі даних користувача на сервері (бекенді). Ці два домени забезпечують повний цикл безпечної взаємодії з користувачем. З огляду на зростання загроз у першому домені, важливо впроваджувати ефективні методи для підвищення безпеки спілкування з чат-ботом. Це особливо актуально для компаній, що працюють із персональними даними, де більшість таких методів є обов'язковими. Далі розглянемо ключові підходи до безпеки спілкування з чат-ботом у веб-інтерфейсах і мобільних додатках.

© Кушнерьов, О. С., Вадько, К. С., & Кузченко, А. О. (2024). Чат-бот на основі нейромережі для надання інформаційної підтримки бізнесу з позиції кібербезпеки. Сучасний захист інформації, 4(60), 131–140. https://doi.org/10.31673/2409-7292.2024.040014.

Перша важлива складова — автентифікація та авторизація. Автентифікація дозволяє підтвердити особу користувача, хоча вона не завжди є обов'язковою. Наприклад, на сайтах електронної комерції автентифікація може бути необов'язковою, тоді як для доступу до банківських даних вона стає необхідною. Автентифікація засвідчує, що користувач має безпечні та дійсні облікові дані, серед яких найпоширенішими є ім'я користувача, електронна пошта, номер телефону, біометричні дані та паролі. Рекомендується впровадження двофакторної автентифікації для підвищення рівня захищеності. Авторизація, у свою чергу, забезпечує доступ до конфіденційних даних лише для авторизованих користувачів, що дозволяє зберігати інформацію від несанкціонованого доступу.

Іншим важливим підходом є наскрізне шифрування (E2EE), яке забезпечує конфіденційність обміну, дозволяючи лише учасникам комунікації читати повідомлення. Кожна сторона має пару ключів — приватний і публічний; наприклад, за допомогою алгоритму RSA повідомлення шифрується відкритим ключем і розшифровується приватним. Захист конфіденційності ключів є критично важливим для забезпечення безпеки. Чат-боти, що не працюють з персональними даними, можуть не використовувати E2EE, проте базовий рівень шифрування забезпечується передачею даних через HTTPS. Вимоги статті 32(a) GDPR зобов'язують компанії шифрувати персональні дані, і хоча популярні платформи, такі як WhatsApp, Facebook Messenger, Telegram, підтримують шифрування, законодавчі вимоги можуть конфліктувати з конфіденційністю, вимагаючи доступ до даних у відкритому вигляді.

Нарешті, для передачі конфіденційної інформації в чат-ботах можуть використовуватися повідомлення, що самознищуються. Це особливо важливо для чат-ботів, пов'язаних із медичними або фінансовими сервісами, оскільки стаття 5(e) GDPR регламентує, що персональні дані не повинні зберігатися довше, ніж це необхідно. У США конфіденційність захищеної медичної інформації (PHI) також підлягає суворому регулюванню, що забезпечує повну конфіденційність і безпеку зберігання [8]. Захищена медична інформація (PHI) відповідно до законодавства США включає дані про стан здоров'я або оплату медичних послуг, що збираються відповідними організаціями. Дотримання вимог GDPR передбачає забезпечення конфіденційності, що означає, що лише наміри користувача можуть бути зареєстровані для аудиту, а особиста інформація ніколи не повинна бути доступною [1].

Висновки

У сучасному цифровому середовищі, де чат-боти та нейронні мережі активно інтегруються в бізнес-процеси, комунікацію та автоматизацію рутинних завдань, постає важливе завдання забезпечення кібербезпеки цих технологій. Чат-боти, які функціонують на основі нейронних мереж, здатні суттєво підвищити ефективність обслуговування, зокрема в обробці запитів, інформуванні клієнтів та підтримці зворотного зв'язку. Проте разом із перевагами ці технології привносять і нові виклики в сфері кібербезпеки, що потребують глибокого аналізу та надійних рішень.

Сучасні кіберзагрози для чат-ботів включають атаки на канали зв'язку, які можуть призвести до перехоплення конфіденційних даних, вразливостей в архітектурі систем та проблем конфіденційності, що загрожують як компаніям, так і їхнім клієнтам. Ненадійний захист може викликати серйозні наслідки, включаючи фінансові втрати, витоки даних та втрату довіри користувачів до системи. Тому впровадження надійних стратегій захисту є обов'язковою умовою для забезпечення безпечного використання чат-ботів і технологій на основі штучного інтелекту.

Одним із пріоритетних завдань для розробників і користувачів цих систем є забезпечення захищеного обміну повідомленнями та захист даних, що передаються через канали зв'язку. Впровадження наскрізного шифрування, багатофакторної автентифікації та регулярного моніторингу системи для виявлення аномальної активності є ключовими кроками для зниження ризиків. Крім того, особливу увагу слід приділити налаштуванню протоколів

конфіденційності, які забезпечуватимуть відповідність нормативним вимогам та захист особистих даних користувачів.

Таким чином, забезпечення кібербезпеки для чат-ботів та систем, що функціонують на основі штучного інтелекту, є не лише технічним завданням, а й питанням довіри з боку користувачів та компаній. Розробники повинні приділяти особливу увагу безпеці під час впровадження цих рішень, забезпечуючи, щоб високий рівень автоматизації та зручність користування не негативно впливали на конфіденційність і цілісність даних. Лише так можна забезпечити надійність та довговічність використання сучасних технологій в умовах цифрової трансформації.

Робота виконана в рамках держбюджетної науково-дослідної роботи № 0124U000550 "Моделювання механізмів протидії організованій та транснаціональній кіберзлочинності у воєнний та післявоєнний часи".

Перелік посилань

1. Абомхара М., Кієн Г. М. Кібербезпека та Інтернет речей: вразливості, загрози, зловмисники та атаки. Журнал кібербезпеки та мобільності . 2015. Вип. 4, № 1. С. 65–88. URL: https://doi.org/10.13052/jcsm2245-1439.414
2. Арора А., Арора А., Макінтайр Дж. Розробка чат-ботів для кібербезпеки: оцінка загроз за допомогою аналізу настроїв у соціальних мережах. Стійкість . 2023. Вип. 15, немає. 17. С. 13178. URL: https://doi.org/10.3390/su151713178
3. Технологія чат-бот як чинник комп'ютерно-посередницької комунікації цифрового суспільства. Humanities Studies. 2022. № 12(89). С. 130–141. URL: https://doi.org/10.26661/hst-2022-12-89-15
4. Новах М. Р., Летвиненко О. В., Виганяйло С. М., Виганяйло С. М., Підготовка правоохоронців в системі МВС України в умовах воєнного стану : зб. наук. пр. / МВС України, Харків. нац. ун-т внутр. справ, Каф. тактич. та спец. фіз. підготовки ф-ту № 3, Наук. парк "Наука та безпека". Харків : ХНУВС, 2022. 440 с. URL: https://univd.edu.ua/science-issue/issue/5612
5. Піскозуб А., Журавчак Д., Толкачова А. Дослідження вразливостей у чатботах з використанням великих мовних моделей. Ukrainian Scientific Journal of Information Security. 2023. Т. 29, № 3. С. 111–117. URL: https://doi.org/10.18372/2225-5036.29.18069
6. Прокопенко Т. О., Обойщик О. Б. Особливості використання чатботів для бізнесу у сучасних месенджер чатах. Вісник Черкаського державного технологічного університету. Серія: Технічні науки. 2019. № 1. С. 11–16. URL: https://doi.org/10.24025/2306-4412.1.2019.165418
7. Стаття 32 GDPR. Безпека опрацювання | GDPR Text, Translation and Commentary. URL: https://gdpr-text.com/uk/read/article-32/ (дата звернення: 04.11.2024).
8. Стаття 5 GDPR. Принципи, які стосуються обробки персональних даних | GDPR Text, Translation and Commentary. URL: https://gdpr-text.com/read/article-5/ (дата звернення: 04.11.2024).
9. How to Make Chatbots Productive – A User-Oriented Implementation Framework / A. Janssen et al. International Journal of Human-Computer Studies. 2022. P. 102921. URL: https://doi.org/10.1016/j.ijhcs.2022.102921
10. Top 10 for Large Language Model Applications | OWASP Foundation. OWASP Foundation, the Open Source Foundation for Application Security | OWASP Foundation. URL: https://owasp.org/www-project-top-10-for-large-language-model-applications/ (дата звернення: 04.11.2024).
11. Pandey S., Sharma S. A comparative study of retrieval-based and generative-based chatbots using Deep Learning and Machine Learning. Healthcare Analytics. 2023. P. 100198. URL: https://doi.org/10.1016/j.health.2023.100198
12. Лаптєв О.А., Собчук В.В., Савченко В.А. Метод підвищення завадостійкості системи виявлення, розпізнавання і локалізації цифрових сигналів в інформаційних системах. Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. К.: ВІКНУ, Вип. 66. 2019. С. 124 – 132.
13. Лаптєв О.А., Савченко В.А. Локалізація засобів негласного отримання інформації методом найменших квадратів. Телекомунікаційні та інформаційні технології: науковий журнал. К.: ДУТ, №4 (65). 2019. С57 – 70.
14. Стефурак О.Р., Тихонов Ю.О., Лаптєв О.А., Зозуля С.А. Удосконалення стохастичної моделі з метою визначення загроз пошкодження або несанкціонованого витоку інформації. Сучасний захист інформації: науково-технічний журнал. К.: ДУТ, 2020. № 2(42)., С 19 – 26.
15. Zamrii I., Sobchuk V., Laptiev O., Savchenko V., Shkapa V., Kovalenko V. and Kotok V. Fractal Functions and Their Application to Source Data Coding. ARPN Journal of Engineering and Applied Sciences. Vol. 17, No. 4, 2022. P. 424 – 435.

Надійшла 06.11.2024