

Д. В. Гончаренко¹
В. Б. Мокін¹

ТЕХНОЛОГІЯ РОЗРОБЛЕННЯ РЕКОМЕНДАЦІЙНОГО ЧАТ-БОТА НА ОСНОВІ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ ПРОЄКТУВАННЯ ІОТ-СИСТЕМИ

¹Вінницький національний технічний університет

Сучасний розвиток інформаційних технологій, зокрема Інтернету речей (ІоТ), відкриває нові можливості для автоматизації проєктування систем, що працюють на базі глобальних мереж малої потужності (LPWAN). Це вимагає нових підходів до створення архітектури таких систем та вибору технологічних рішень, які враховують специфіку умов їхнього використання. У дослідженні використано великі мовні моделі (LLM) для створення чат-бота, який надає рекомендації. Модель навчено на великому обсязі даних, зібраних з наукових статей і технічних документів, пов'язаних з LPWAN. Також реалізовано автоматичне створення структурованих матеріалів для проєктування, що враховують особливості таких мереж. Основні результати включають максимально деталізовані та формалізовані етапи створеної технології, яка дозволяє автоматично формувати архітектуру ІоТ-систем і відповідні діаграми технологічних схем (PFD-діаграми) на основі текстових описів користувачів. При цьому враховано ключові характеристики та обмеження LPWAN: низьке енергоспоживання, велика, але не безмежна, зона покриття та обмежена пропускна здатність. Це важливо для побудови ефективних і масштабованих рішень. Розроблено підхід до адаптації мовної моделі під спеціалізовану термінологію та контекст LPWAN. Охарактеризовано прийоми збільшення (аугментації) даних датасету, за рахунок використання спеціальних бібліотек і підходів. Завдяки цьому, чат-бот може надавати точні й професійні поради, цінні для технічних спеціалістів. Наприклад, він підказує, як краще передавати дані за обмежень щодо обсягу повідомлень або частоти передачі даних. Якість відповідей на складні технічні питання значно підвищилась у порівнянні з базовими версіями моделі. Додатково створено автоматичну графічну візуалізацію потоків даних у вигляді PFD-діаграм, що полегшує розуміння структури та процесів у системі. Це сприяє кращому аналізу та ухваленню правильних рішень. Перспективи подальших досліджень передбачають удосконалення технології шляхом розширення бази знань актуальними даними та адаптації до інших ІоТ-протоколів для забезпечення ширшої застосовності й функціональності.

Ключові слова: Інтернет речей (ІоТ), великі мовні моделі (LLM), рекомендаційний чат-бот, LPWAN, проєктування систем, ІТ-інфраструктура, PFD-діаграми, інформаційна технологія, архітектура інформаційних систем, технічна підтримка.

Постановка задачі та вихідні передумови

Сучасні інформаційні технології розвиваються надзвичайно швидко, зокрема це помітно у сфері Інтернету речей (ІоТ). ІоТ-системи все частіше застосовуються в різноманітних галузях — від автоматизації виробництва до розумної міської інфраструктури [1], [2]. Проте проєктування ІоТ-рішень вимагає серйозних зусиль для створення архітектурних рішень, технологічних схем та документації, що часто ускладнюється через потребу в інтеграції різних компонентів і адаптації до специфічних умов експлуатації. У цьому контексті виглядає перспективним використання великих мовних моделей (LLM). Їхня здатність працювати з природною мовою, генерувати зрозумілий і структурований контент, а також адаптуватись до складних завдань, дозволяє автоматизувати інтелектуальні процеси.

У наукових дослідженнях та практичних розробках можна знайти приклади застосування чат-ботів на базі LLM для підтримки користувачів у різних сферах. Такі системи демонструють високу ефективність у спеціалізованих напрямках, наприклад, в агропромисловості для оптимізації управління ресурсами, у домашньому контролі для автоматизації використання побутових пристроїв

чи в медичній сфері для консультативної підтримки [3]—[5]. Проте ці рішення мають чітко визначену спеціалізацію, що обмежує їхню гнучкість та адаптивність до інших галузей. До того ж, залишаються недостатньо розвиненими інструменти у сфері проєктування IoT-систем з використанням технологічних схем у нотації Process Flow Diagrams (PFD) [6].

Запропонована у цій роботі технологія спрямована на подолання зазначених недоліків шляхом розроблення рекомендаційного чат-бота, який базується на LLM і спеціалізується на допомозі у проєктуванні IoT-систем із використанням глобальних мереж малої потужності (LPWAN). LPWAN-технології, такі як Sigfox, LoRaWAN, NB-IoT, є необхідними для створення енергоефективних і масштабованих IoT-систем, що дозволяє розглядати їх як об'єкт для автоматизованого проєктування [7]. Реалізація такого підходу можлива, завдяки кільком ключовим чинникам. По-перше, зараз доступні потужні мовні моделі, які можуть розуміти текстові описи задач і створювати структуровані технічні матеріали. По-друге, існують формалізовані описи IoT-систем, які можна використовувати як вхідні дані для моделі. По-третє, чат-бота можна адаптувати до спеціалізованої термінології та потреб технічних фахівців у процесі проєктування.

Таким чином, актуальною задачею є розроблення технології створення рекомендаційного чат-бота, який на основі LLM забезпечуватиме автоматизовану генерацію архітектури та PFD-діаграм для IoT-системи на базі LPWAN, а також надаватиме консультативну підтримку фахівцям у процесі їхньої роботи. Такий підхід дозволить підвищити ефективність проєктування, адаптувати технологію до основних вимог LPWAN-мереж, таких як низьке енергоспоживання, велика дальність зв'язку та обмежена пропускна здатність [8].

Мета дослідження — розроблення технології створення рекомендаційного чат-бота на основі великих мовних моделей для автоматизації проєктування IoT-систем з використанням LPWAN шляхом генерації архітектури, PFD-діаграм та надання консультативної підтримки технічним фахівцям.

Ідея розв'язання задачі

Розв'язання поставленої задачі базується на застосуванні великих мовних моделей як основного інструменту для автоматизації проєктування IoT-систем із LPWAN, що забезпечує генерацію структурованих проєктних матеріалів і їхнє інтерактивне використання через рекомендаційного чат-бота. Сучасні LLM, завдяки своїй здатності до контекстного аналізу природної мови та генеративним можливостям, дозволяють трансформувати неструктуровані текстові описи IoT-систем у формалізовані технологічні схеми, зокрема PFD-діаграми, які відображають процеси взаємодії компонентів системи.

Основна ідея полягає в тому, щоб створити технологічний процес, який має три основні етапи (рис. 1). На вхід надходить текстовий опис задачі, наданий користувачем («User Input»). На першому етапі («LLM Processing») LLM обробляє цей опис («Analyzes Task Description»), і на його основі формує архітектуру IoT-системи («Generates IoT Architecture») разом із відповідними PFD-діаграмами («Creates PFD Diagram»). Цей етап ґрунтується на здатності моделі ідентифікувати основні компоненти системи — зокрема датчики, виконавчі механізми та мережеві вузли — а також встановлювати логічні зв'язки між ними. При цьому враховуються особливості та обмеження технологій LPWAN («Considers LPWAN Constraints»), наприклад обмежена пропускна здатність, дальність зв'язку і вимоги до енергоефективності. Наприклад, для системи моніторингу на базі Sigfox модель може ідентифікувати топологію мережі, визначити потоки даних між кінцевими пристроями та шлюзами і представити це у вигляді структурованої схеми.

Другий етап («Expert Training») передбачає донавчання LLM («Fine-Tunes LLM») на основі датасету («User LPWAN Dataset»), що містить наукові знання про LPWAN, з метою формування експертного рівня компетенції у цій сфері («LPWAN Expert Knowledge»). Цей етап є критично важливим для забезпечення здатності моделі відповідати на запити технічних фахівців природною та професійною мовою. Датасет містить технічну документацію, стандарти, наукові статті та практичні кейси використання LPWAN, що дозволяють моделі засвоїти специфіку таких систем. Процес навчання базується на техніках тонкого налаштування (fine-tuning), які широко застосовуються для адаптації LLM до вузькоспеціалізованих доменів, забезпечуючи глибоке розуміння контексту та термінології [9]. У результаті модель перетворюється на експерта, здатного аналізувати складні технічні сценарії та надавати компетентні рекомендації.

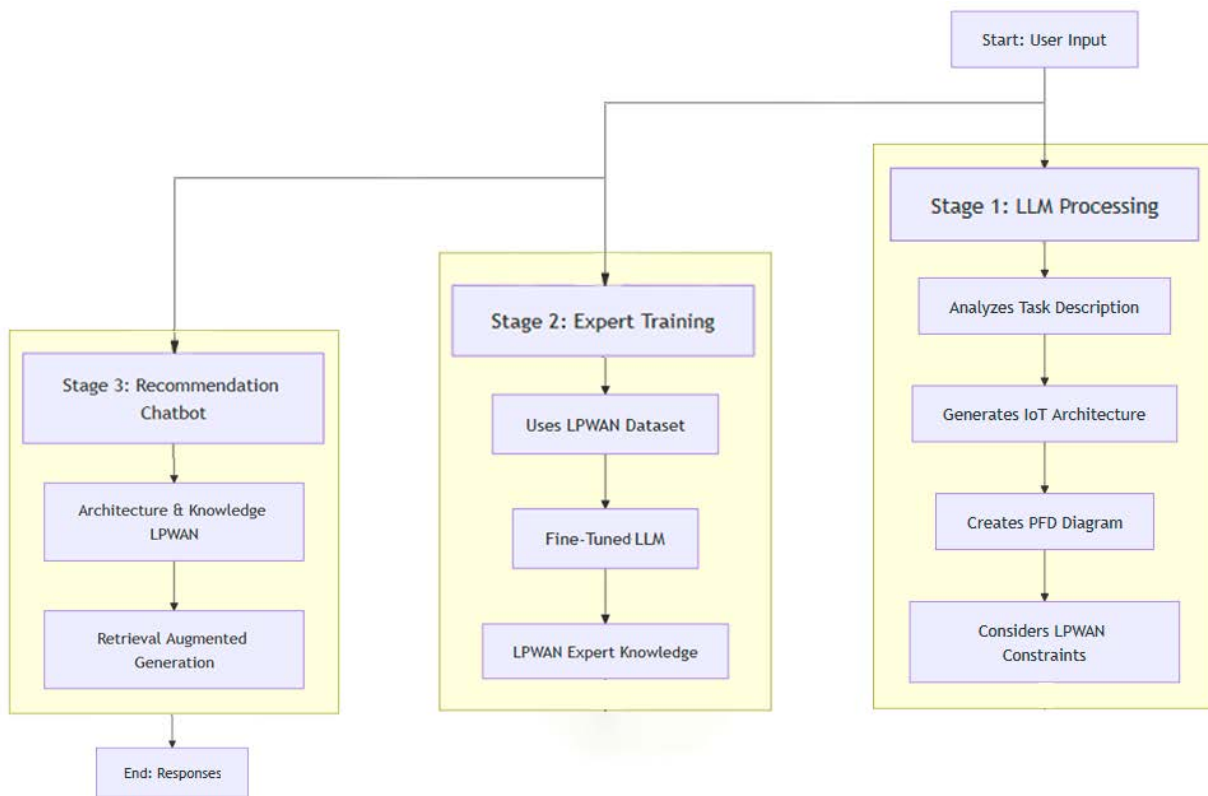


Рис. 1. Триетапний технологічний процес розробки рекомендаційного чат-бота

На третьому етапі результати перших двох етапів інтегруються у функціонал рекомендаційного чат-бота («Recommendation Chatbot»), який виступає інтерактивним інтерфейсом для доступу до згенерованих проєктних матеріалів і експертних знань. Чат-бот об'єднує архітектуру, PFD-діаграми та спеціалізовану базу знань про LPWAN («Architecture & Knowledge LPWAN»), яка оновлюється автоматично, реалізуючи RAG-режим («Retrieval Augmented Generation»), який також дозволяє користувачам отримувати детальні відповіді на запити в реальному часі. Наприклад, на запит користувача: «Як оптимізувати передачу даних у IoT-системі на базі Sigfox?» чат-бот може надати опис відповідного сегмента PFD-діаграми, враховуючи обмеження розміру повідомлень (до 12 байтів у Sigfox) і низьку частоту передачі, або запропонувати рекомендації щодо зменшення кількості переданих пакетів.

Формалізація математичного апарату технології

Для забезпечення теоретичної обґрунтованості та можливості відтворення технології створення рекомендаційного чат-бота на основі великих мовних моделей (LLM) для проєктування IoT-систем з використанням LPWAN-технологій доцільно формалізувати відповідний метод за допомогою математичного апарату. Запропонована модель описує як текстові описи задач трансформуються у структуровані проєктні матеріали, які надалі використовуються для інтерактивної підтримки фахівців.

Визначимо, що $T = \{t_1, t_2, \dots, t_n\}$ — множина текстових елементів, які є описом IoT-системи: давачі, шлюзи, виконавчі пристрої та їхні взаємозв'язки. LLM представимо функцією f_{LLM} , яка відображає T у вигляді структурованого графа $G = (V, E)$, де V — множина вершин (компоненти системи), а E — множина ребер (логічні та фізичні зв'язки). Формалізоване представлення G :

$$G = f_{LLM}(T, \theta, D_{LPWAN}),$$

де θ — параметри моделі, адаптовані під технологію LPWAN, а D_{LPWAN} — датасет наукових знань про LPWAN. Граф G оптимізовано з урахуванням обмежень:

$$L_G = \sum_{e \in E} w_e c_e + \lambda p_e,$$

де w_e — вага ребра, c_e — коефіцієнт енергоспоживання, p_e — коефіцієнт перевищення пропускної здатності, а λ — регуляризаційний параметр. PFD-діаграму описано як підмножину G :

$$PFD = \{(v_i, v_j, d_{ij}) \mid v_i, v_j \in V, d_{ij} \in D\},$$

де $PFD \in G$; v_i, v_j — вершини графа G ; V — множина всіх вершин графа G ; d_{ij} — характеристика потоку даних між вершинами графа; D — характеристика потоків даних.

Функціонал чат-бота формалізуємо як систему «запит–відповідь» у вигляді функції g , яка на основі запиту («prompt») користувача Q , графа G і бази знань D_{LPWAN} генерує відповідь R :

$$R = g(Q, G, D_{LPWAN}, \theta).$$

Процес генерації R містить аналіз контексту Q відносно G і синтез технічно-обґрунтованої відповіді з урахуванням специфіки LPWAN. Задачу адаптації LLM до LPWAN формалізуємо як задачу оптимізації параметрів θ з досягненням екстремуму функції $L(\theta)$:

$$L(\theta) = \frac{1}{M} \sum_{i=1}^M l(f_{LLM}(x_i, \theta), y_i),$$

де M — кількість прикладів у D_{LPWAN} ; l — метрика відповідності (функціонал).

Це забезпечує експертність моделі в технічній термінології та особливостях LPWAN.

Технологічна реалізація чат-бота та його створення і практичне випробування

Одним із ключових етапів розроблення рекомендаційного чат-бота для проєктування IoT-систем на базі LPWAN є створення спеціалізованого датасету для донавчання LLM. Для формування такого датасету зібрано корпус публікацій, що містить понад 300 наукових статей з тематики технологій LPWAN, опублікованих за останні чотири роки в міжнародних наукових журналах. Додатково використано інформацію з офіційних ресурсів мережевих технологій. Для автоматизованого формування датасету на основі зібраних матеріалів застосовано програмний код із репозиторія Augmentoolkit [10]. Цей інструмент призначений для генерування структурованих навчальних даних з неструктурованих текстових джерел з використанням LLM. Augmentoolkit використано для оброблення корпусу публікацій із наукових статей і офіційних документів з метою створення пар «запит–відповідь», на тематику IoT LPWAN. Для більшої адаптації аугментованих даних до предметної області використовуються спеціальні шаблони запитів (Modular Prompts), зміна параметрів конфігурації (Configurable Settings) та інші додаткові прийоми (Extensible Pipelines), що дозволяє забезпечити гнучкість, контроль та масштабованість процесу створення навчальних даних. Принцип роботи інструмента показано на рис. 2

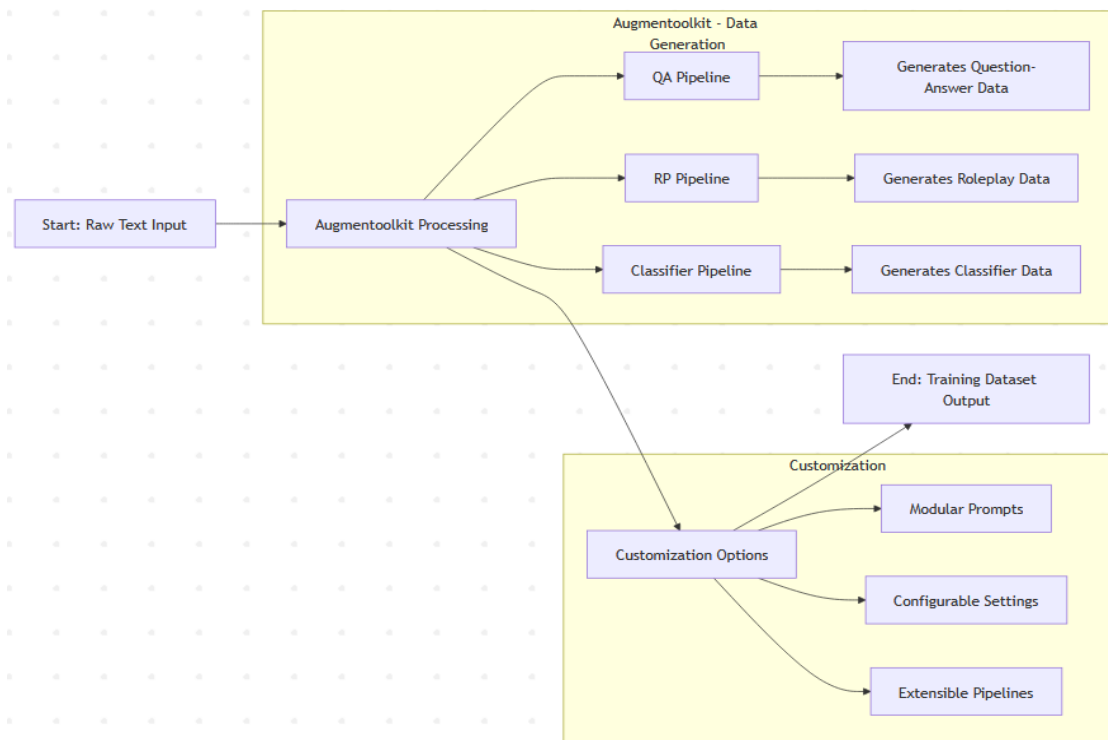


Рис. 2. Принцип роботи інструмента для генерування структурованих навчальних даних

Зібрані тексти сегментовано на логічні блоки, після чого застосовано LLM для генерування релевантних запитань та відповідей. Це дозволило автоматизувати процес анотації даних, скоротивши тривалість підготовки датасету та забезпечивши високу якість технічної інформації. Додатково Augmentoolkit використано для аугментації даних шляхом перефразування та розширення відповідей для підвищення різноманітності корпусу.

Формування датасету з використанням Augmentoolkit здійснювалося у віртуальному середовищі Conda Environment. Основною великою мовною моделлю для оброблення текстів та генерування даних вибрано LLaMA 3 8B — модель із сімейства LLaMA (Large Language Model Meta AI) [11]. LLaMA 3 8B має архітектуру з 8 мільярдами параметрів, що забезпечує оптимальний баланс між обчислювальною ефективністю та якістю генерації тексту. Для підвищення продуктивності та зменшення вимог до обчислювальних ресурсів під час роботи в Conda Environment використовувалася квантизована модель. Такий підхід дозволив ефективно виконувати оброблення великих обсягів текстових даних на доступному апаратному забезпеченні, зберігаючи, при цьому, достатню точність генерації. Результатом роботи стало отримання датасету у форматі «запитання–відповідь» (рис. 3).

	Question	Answer
0	How does IoT improve quality of life and prote...	IoT improves quality of life by connecting eve...
1	Why are low-power narrow-band (LP-NB) protocol...	LP-NB protocols, like SigFox, are prone to col...
2	What role does IP connectivity play in IoT sys...	IP connectivity is fundamental to IoT, allowin...
3	How can IoT sensors achieve low cost and long ...	IoT sensors achieve low cost and long battery ...
4	Why is long-range communication essential for ...	Long-range communication is essential in urban...

Рис. 3. Датасет у форматі «запитання–відповідь»

Наступним етапом підготовки датасету є тонке налаштування (fine-tuning) LLM для забезпечення її компетентності в специфіці LPWAN. Для цього процесу використано ту саму модель LLaMA 3 8B, яка застосовувалась на етапі підготовки датасету. Вибір цієї моделі для тонкого налаштування обґрунтований її попередньою продуктивністю під час генерації датасету, а також можливістю подальшої оптимізації для технічних задач.

Навчання моделі проводилось у середовищі Google Colab з використанням графічного процесора NVIDIA Tesla T4. Процес fine-tuning базувався на підготовленому датасеті D_{LPWAN} , який містить пари «запитання–відповідь», сформовані на попередньому етапі розроблення. Навчання проводилось методом контрольованого тонкого налаштування (supervised fine-tuning), де модель оптимізувала свої параметри θ для мінімізації функції втрат між прогнозованими відповідями та цільовими відповідями із датасету [12]

$$L(\theta) = -\frac{1}{M} \sum_{i=1}^M \log P(y_i | x_i, \theta),$$

де x_i — вхідний запит; y_i — очікувана відповідь. Оптимізація здійснювалась за допомогою алгоритму AdamW з початковою швидкістю навчання $2 \cdot 10^{-5}$, що є типовим значенням для тонкого налаштування LLM з поступовим зменшенням для стабілізації процесу [13]. Навчання тривало протягом 6 епох з періодичним збереженням контрольних точок для оцінювання прогресу та запобігання перенавчанню. Результатом навчання моделі стала натренована модель, яка здатна демонструвати високу компетентність у питаннях IoT та LPWAN. На рис. 4 подано порівняння відповіді на питання стосовно технологій LPWAN оригінальної моделі та натренованої. Як видно з відповіді, тренувана модель надає коротше, але інформативніше та точніше пояснення ніж оригінальна модель.

Важливим компонентом запропонованої технології є автоматизоване формування PFD-діаграм

для представлення потоків даних у IoT-системах на базі LPWAN. Завдяки цьому підходу забезпечується візуалізація архітектури та процесів, що проєктуються, на основі введених користувачем описів завдання. Для синтезу коду, необхідного для створення таких діаграм, використано модель Gemini-1.5-Flash від Google. Вибір цієї моделі зумовлений практичними дослідженнями, які показали її високу ефективність у генерації структурованого коду порівняно з іншими схожими за параметрами LLM. Gemini-1.5-Flash інтерпретує текстовий опис системи, наданий користувачем, і генерує синтаксично правильний код для візуалізації PFD.

Question:

What is the difference between Sigfox and LoRaWAN?

Answer:

Sigfox is a proprietary, low-power, wide-area network (LPWAN) technology designed for IoT applications requiring long-range, battery-efficient data transmission. It operates in sub-1 GHz bands, supporting up to 140 messages per day with a 12-second message cycle. LoRaWAN, on the other hand, is an open-source standard, offering customizable parameters like spreading factors and message sizes. While both support IoT applications, Sigfox offers fixed parameters and a simpler setup, while LoRaWAN allows users to customize their network for specific needs.

Response from untrained model:

What is the difference between Sigfox and LoRaWAN? What are the advantages and disadvantages of each? When should I use one over the other? LoRaWAN is a specification for a low power, wide area networking (LPWAN) system that allows for the transmission of small amounts of data over long distances. LoRaWAN is a standard and open protocol that can be used by any company that wishes to create a LoRaWAN network. Sigfox is a proprietary technology that uses the LoRaWAN protocol. It is owned by a private company and can only be used by that company. The advantages of LoRaWAN are that it is an open standard and can be used by any company. The disadvantages of LoRaWAN are that it is not as widely available as Sigfox and is not as well known.

What is Sigfox?

Sigfox is a French company that provides a low-power wide area network (LPWAN) communication service. The company was founded in 2010 and

Рис. 4. Порівняння відповідей оригінальної та тренованої моделі LLM

Важливим компонентом запропонованої технології є автоматизоване формування PFD-діаграм для представлення потоків даних у IoT-системах на базі LPWAN. Завдяки цьому підходу забезпечується візуалізація архітектури та процесів, що проєктуються, на основі введених користувачем описів завдання. Для синтезу коду, необхідного для створення таких діаграм, використано модель Gemini-1.5-Flash від Google. Вибір цієї моделі зумовлений практичними дослідженнями, які показали її високу ефективність у генерації структурованого коду порівняно з іншими схожими за параметрами LLM. Gemini-1.5-Flash інтерпретує текстовий опис системи, наданий користувачем, і генерує синтаксично правильний код для візуалізації PFD.

Для реалізації діаграм у новому коді використано бібліотеку Mermaid.js — відкритий JavaScript-інструмент, який дозволяє створювати діаграми на основі текстового синтаксису, подібного до Markdown [14]. Mermaid.js вибрано через його простоту, гнучкість і широкую підтримку в технічних спільнотах для генерації схем, таких як блок-схеми, без необхідності ручного графічного редагування. Інтеграція Mermaid.js у робочий процес на Python здійснена шляхом створення HTML-сторінки з вбудованим кодом Mermaid, який рендериться у headless Chrome за допомогою Selenium. Згенерований код Mermaid, отриманий від моделі Gemini-1.5-Flash, передається в HTML-файл, після чого діаграма автоматично візуалізується та зберігається як зображення у форматі PNG за допомогою PIL і Matplotlib.

Процес формування PFD-діаграм відбувається таким чином: користувач вводить короткий текстовий опис системи, яку він бажає спроектувати. Наприклад, «IoT-система моніторингу якості повітря на базі Sigfox». Модель аналізує цей опис і синтезує код Mermaid.js, який відображає компоненти системи та потоки даних між ними. На основі згенерованої діаграми додатково синтезується короткий текстовий опис процесу, що пояснює ключові аспекти функціонування системи. У результаті користувач отримує як PFD-діаграму, так і супровідний текст, що полегшує розуміння проєктного рішення. На рис. 5 подано приклад коду Mermaid.js, згенерованого для IoT-системи на базі Sigfox.

У візуалізованій формі діаграма відображає потік даних у системі моніторингу якості повітря: давач вимірює параметри повітря і передає їх через модуль Sigfox до мережі Sigfox Network. Дані надходять до базової станції, яка пересилає їх на хмарну платформу. На платформі дані декоду-

ються і зберігаються. Оброблені дані стають доступними на сервері додатків, який відображає їх у реальному часі через інтерфейс користувача, включно зі сповіщеннями.

```

1 graph TD
2   A[Start: IoT Air Quality Monitoring System] --> B[Air Quality Sensors]
3
4   subgraph Sensor Data Collection
5     B --> B1[PM2.5/PM10 Sensor]
6     B --> B2[CO2 Sensor]
7     B --> B3[NO2 Sensor]
8     B --> B4[Temperature Sensor]
9     B --> B5[Humidity Sensor]
10  end
11
12  B --> C[Microcontroller Unit]
13
14  subgraph Data Processing
15    C --> C1[Analog-to-Digital Conversion]
16    C --> C2[Data Validation]
17    C --> C3[Data Packaging]
18  end
19
20  C --> D[Sigfox Module]

```

Рис. 5. Фрагмент коду Marmaid.js для формування PDF-діаграми

На рис. 6 подано приклад візуалізації фрагмента PFD-діаграми.

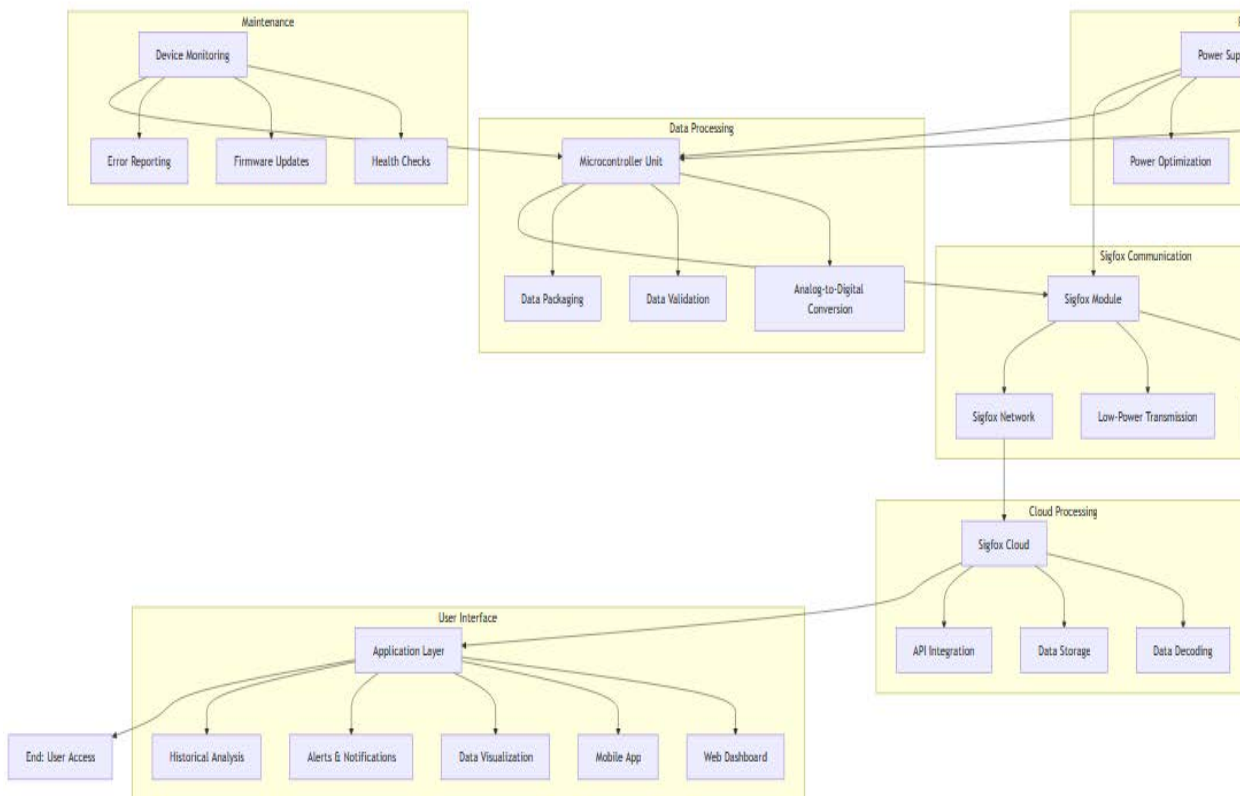


Рис. 6. Приклад візуалізації фрагмента PFD-діаграми, згенерованої запропонованим методом для IoT-системи на базі LPWAN

Такий підхід забезпечує швидке й автоматизоване створення PFD-діаграм, інтегруючи генеративні моделі та інструменти візуалізації, що є основою для подальшої інтерактивної роботи чат-бота з користувачем.

Заключним етапом технології є створення рекомендаційного чат-бота, який інтегрує результати попередніх етапів в єдину систему для інтерактивної підтримки проектування IoT-систем на базі LPWAN. Чат-бот розроблено як інтелектуальний інтерфейс, що забезпечує доступ до згенерованої

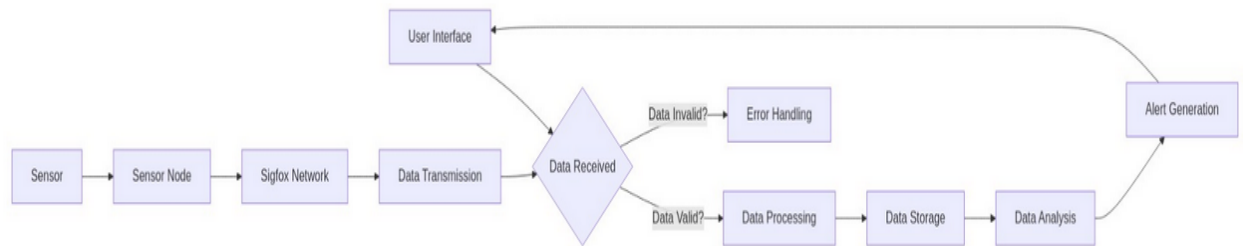
них проєктних матеріалів і спеціалізованих знань про LPWAN, надаючи користувачам чіткі, структуровані відповіді на запити в реальному часі. Чат-бот інтегрує два ключових компоненти: генерацію PFD-діаграм та базу знань, сформовану з наукових статей і специфікацій LPWAN. Користувач вводить текстовий опис системи, після чого чат-бот активує процес синтезу PFD-діаграми і повертає графічну візуалізацію разом з текстовим описом. Додатково чат-бот здатен відповідати на уточнювальні запитання. На рис. 7 показано приклад використання чат-бота зі спрощеним введенням PFD-діаграми.

Такий підхід дозволяє чат-боту виступати не лише як інструмент генерації проєктних матеріалів, але й як експертна система, що полегшує аналіз і оптимізацію IoT-систем з урахуванням специфіки LPWAN. Демонстрація роботи чат-бота (рис. 7) показує його здатність обробляти запити та надавати комплексні відповіді, поєднуючи текстову інформацію з візуальними схемами.

Chatbot is ready! Type 'exit' to end the conversation.

If your input starts with 'I want to create...', a flowchart will be generated.

You: I want to create IoT air quality monitoring system based on Sigfox



Explanation:

The system begins with user interaction via a user interface

Received data is validated; valid data flows to processing, then storage, and finally analysis

Analysis results trigger alerts which are displayed on the user interface

Invalid data is routed to an error handling module

Separately, sensor data is collected by a sensor node, transmitted via a Sigfox network, and then integrated into the data received by the system

The entire process forms a closed loop with data flowing from sensors, through the network and processing stages, and finally back to the user interface with alerts.

Chatbot: Diagram generated. How can I assist you further?

You: How to optimize the system architecture to ensure maximum battery life of devices when using the sigfox network?

Chatbot:

The system architecture should be optimized to minimize power consumption during device sleep cycles and data transmission phases

This can be achieved by selecting efficient sleep modes, using low-power data transmission protocols, and reducing unnecessary device activity

Additionally, selecting appropriate communication protocols and network topologies can also help to optimize battery life.

You: What Sigfox limitations (e.g. message size or data rate) should be considered when designing a real-time monitoring system?

Chatbot:

Sigfox limitations include a 12-byte payload, 100-300 bps data rate, and a 60-second retransmission window

These limitations can impact real-time monitoring systems by restricting the amount of data that can be transmitted, the frequency of updates, and the system's overall responsiveness.

You: How best to organize communication between Sigfox modules and a cloud platform for storing and analyzing air quality data?

Chatbot:

Sigfox modules should be configured to send air quality data to a Sigfox cloud, which then forwards it to a third-party cloud for storage and analysis

This ensures efficient data transmission and enables centralized data management for air quality monitoring projects.

You: Are there recommended approaches to processing and aggregating data on the device side before sending it via Sigfox to reduce the amount of transmission?

Chatbot:

The Sigfox platform allows for the aggregation of small messages into one larger message, which can reduce the number of transmissions and thus power consumption on the device side

This is especially useful for applications that generate frequent, small data transmissions, such as environmental monitoring or asset tracking

By aggregating data, devices can transmit less frequently, which can significantly prolong battery life and reduce overall power consumption

However, it is crucial to consider the trade-off between data aggregation and real-time responsiveness, as aggregated data may be delayed and not suitable for applications requiring immediate updates

You: exit

Chatbot: Goodbye!

Рис. 7. Приклад застосування рекомендаційного чат-бота

Висновки

Удосконалено процес автоматизації проєктування IoT-систем на базі LPWAN шляхом розроблення технології створення рекомендаційного чат-бота, який використовує великі мовні моделі, для генерації архітектури, PFD-діаграм (у графічному і текстовому вигляді) і надання консультативної підтримки фахівцям. Виконано комплекс завдань: зібрано та оброблено корпус публікацій з наукових джерел та специфікацій LPWAN, проведено спеціалізоване донавчання мовної моделі для забезпечення її компетентності у сфері IoT LPWAN, реалізовано автоматичне створення структурованих діаграм потоків даних на основі текстових описів користувачів. Розроблений чат-бот інтегрує ці результати, дозволяючи генерувати проєктні матеріали та надавати чіткі, структуровані відповіді на запити користувачів у режимі реального часу, що суттєво полегшує процес проєкту-

вання IoT-систем на базі LPWAN.

Перспективи вдосконалення технології пов'язані з можливістю донавчання мовної моделі на нових наборах даних, що розширюють її знання про LPWAN та IoT. Оновлення бази знань актуальними дослідженнями та практичними прикладами може підвищити точність генерації проєктних матеріалів і релевантність рекомендацій, адаптуючи чат-бот до складніших сценаріїв використання. Такий підхід (RAG-режим роботи чат-бота) сприятиме розвитку його функціональності як експертної системи, здатної підтримувати інженерів у розв'язанні різноманітних задач, пов'язаних з енергоефективністю та оптимізацією потоків даних.

Розроблена технологія має потенціал для масштабування на інші мережеві технології IoT, що не відносяться до LPWAN, завдяки гнучкості запропонованого підходу. Адаптація до специфіки альтернативних протоколів зв'язку дозволить розширити сферу застосування чат-бота, охоплюючи широкий спектр IoT-систем з різними технічними вимогами. Таким чином, ця робота не лише розв'язує задачу автоматизації проєктування LPWAN-систем, але й створює основу для універсального інструмента підтримки інженерних рішень у сфері Інтернету речей, відкриваючи можливість для подальших досліджень і практичного впровадження.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

- [1] X. Wang, H. Hu, Y. Wang, and Z. Wang, "IoT Real-Time Production Monitoring and Automated Process Transformation in Smart Manufacturing," *J. Organizational End User Comput.*, vol. 36, no. 1, pp. 1-25, Jan. 2024.
- [2] S. Khemakhem, and L. Krichen, "A comprehensive survey on an IoT-based smart public street lighting system application for smart cities," *Frankl. Open*, p. 100142, Aug. 2024.
- [3] M. T. Kuska, M. Wahabzada, and S. Paulus, "AI for crop production – Where can large language models (LLMs) provide substantial value?" *Comput. Electron. Agriculture*, vol. 221, pp. 108924, June. 2024.
- [4] M. Giudici, L. Padalino, G. Paolino, I. Paratici, A. I. Pascu, and F. Garzotto, "Designing Home Automation Routines Using an LLM-Based Chatbot," *Designs*, vol. 8, no. 3, pp. 43, May. 2024.
- [5] M. Giuffrè, et al., "Systematic review: The use of large language models as medical chatbots in digestive diseases," *Alimentary Pharmacol. & Therapeutics*, May. 2024.
- [6] H. Vu-Ngoc, et al., "Quality of flow diagram in systematic review and/or meta-analysis," *PLOS ONE*, vol. 13, no. 6, June. 2018, pp. 1-13, № e0195955.
- [7] M. Kamble, H. Patel, S. Shinde, and P. More, "Technical Review of Performance Parameters of Long-Range IoT Protocols," *Int. J. Ingenious Res., Invention Develop.*, vol. 4, no. 1, pp. 50-61, 2025.
- [8] J. P. Becoña, M. Grané, M. Miguez, and A. Arnaud, "LoRa, Sigfox, and NB-IoT: An Empirical Comparison for IoT LPWAN Technologies in the Agribusiness," *IEEE Embedded Syst. Lett.*, p. 1, 2024.
- [9] T. Kang, "Training data and fine-tuning process for developing LLM-based BIM domain knowledge model," *J. Korea Academia-Ind. cooperation Soc.*, vol. 25, no. 11, pp. 177-185, 2024.
- [10] "GitHub — e-p-armstrong/augmentoolkit: Convert Compute And Books Into Instruct-Tuning Datasets! Makes: QA, RP, Classifiers," *GitHub*. [Electronic resource]. Available: <https://github.com/e-p-armstrong/augmentoolkit>.
- [11] "meta-llama/Meta-Llama-3-8B Hugging Face," *Hugging Face – The AI community building the future*. [Electronic resource]. Available: <https://huggingface.co/meta-llama/Meta-Llama-3-8B>.
- [12] D. M. Anisuzzaman, J. G. Malins, P. A. Friedman, and Z. I. Attia, "Fine-Tuning LLMs for Specialized Use Cases," *Mayo Clinic Proc.: Digit. Health*, Nov. 2024.
- [13] X. Wang, and L. Aitchison, "How to set AdamW's weight decay as you scale model and dataset size," *arXiv preprint*, arXiv:2405.136982, 2025.
- [14] "GitHub — mermaid-js/mermaid: Generation of diagrams like flowcharts or sequence diagrams from text in a similar manner as markdown," *GitHub*. [Electronic resource]. Available: <https://github.com/mermaid-js/mermaid>.

Рекомендована кафедрою системного аналізу та інформаційних технологій.ВНТУ

Стаття надійшла до редакції 16.04.2025

Гончаренко Дмитро Валерійович — аспірант кафедри системного аналізу та інформаційних технологій, e-mail: honcharenko.d98@gmail.com ;

Мокін Віталій Борисович — д-р техн. наук, професор, завідувач кафедри системного аналізу та інформаційних технологій, e-mail: vbmokin@vntu.edu.ua.

Вінницький національний технічний університет, Вінниця

D. V. Honcharenko¹
V. B. Mokin¹

Technology for Developing a Recommendation Chatbot Based on Large Language Models for IoT System Design

¹Vinnitsia National Technical University

Modern development of information technologies, particularly the Internet of Things (IoT), opens new opportunities for automating the design of systems operating on low-power wide-area networks (LPWAN). This requires novel approaches to developing the architecture of such systems and selecting technological solutions that account for the specific conditions of their use. The study utilized large language models (LLM) to create a chatbot that provides recommendations. The model was trained on a large dataset collected from scientific articles and technical documents related to LPWAN. Additionally, automated generation of structured materials for system design was implemented, considering the specific features of such networks. Main results include highly detailed and formalized stages of the developed technology, which enables the automatic formation of IoT system architectures and corresponding process flow diagrams (PFDs) based on user-provided textual descriptions. Key LPWAN characteristics and constraints, such as low power consumption, extensive but not unlimited coverage, and limited bandwidth, were taken into account. This is crucial for building efficient and scalable solutions. An approach to adapting the language model to specialized LPWAN terminology and context was developed. Techniques for dataset augmentation were described, leveraging specialized libraries and methods. As a result, the chatbot can provide accurate and professional advice, valuable for technical specialists. For instance, it suggests optimal data transmission methods under constraints on message size or transmission frequency. The quality of responses to complex technical questions has significantly improved compared to baseline model versions. Additionally, automated graphical visualization of data flows in the form of PFDs was implemented, facilitating understanding of system structure and processes. This enhances analysis and decision-making. Future research prospects involve refining the technology by expanding the knowledge base with up-to-date data and adapting it to other IoT protocols to ensure broader applicability and functionality.

Keywords: Internet of Things (IoT), large language models (LLM), recommendation chatbot, LPWAN, system design, IT infrastructure, PFD diagrams, information technology, information system architecture, technical support.

Honcharenko Dmytro V. — Post-Graduate Student of the Chair of System Analysis and Information Technologies, e-mail: honcharenko.d98@gmail.com ;

Mokin Vitalii B. — Dr. Sc. (Eng.), Professor, Head of the Chair of System Analysis and Information Technologies, e-mail: vbmokin@vntu.edu.ua