

СЕМАНТИЧНА МАРШРУТИЗАЦІЯ ЗАПИТІВ У ШЛЮЗІ API СИСТЕМИ ТРЕКІНГУ ЗАМОВЛЕНЬ

Безуглий Н.С.

e-mail: nazar.bezuhlyi@nure.ua

Науковий керівник – к.т.н., доц. Імангулова З.А.

Харківський національний університет радіоелектроніки, каф. КМІТ
м. Харків, Україна

This paper presents an LLM-based approach to enhance request routing in a microservice order-tracking system. An API Gateway is extended with semantic understanding to interpret natural-language queries and select the required services. The method converts user intents into a controlled sequence of tool calls with validated parameters. A layered design (security–intelligence–execution) ensures access control, observability, and predictable behavior. Session context supports multi-turn clarification when inputs are incomplete. The approach reduces client-side orchestration and speeds up adding new user scenarios. It is applicable to logistics and other microservice enterprise systems where flexible user interaction is needed.

Сучасні інформаційні системи трекінгу замовлень працюють у середовищі високої динаміки: змінюються вимоги бізнесу, логістичні процеси, канали взаємодії з клієнтами та інтеграції із зовнішніми службами доставки і платіжними рішеннями [1]. У таких умовах мікросервісна архітектура є практичним вибором, оскільки дозволяє незалежно розвивати компоненти системи, масштабувати найкритичніші ділянки та зменшувати ризики під час розгортань. Водночас поділ функціональності між багатьма сервісами ускладнює виконання користувацьких сценаріїв: один запит часто потребує послідовних викликів кількох мікросервісів, узгодження параметрів, збирання часткових відповідей і коректної обробки виняткових ситуацій. Для прикладу, запит користувача «де моє замовлення?» може вимагати знаходження замовлення, отримання останнього статусу, перевірки даних перевізника, уточнення точки видачі або адреси доставки, а також формування зрозумілого підсумку з часовими оцінками. Якщо ж користувач запитує «зміни адресу доставки і скажи, чи встигне кур'єр», потрібно не лише виконати зміну, а й перевірити допустимість дії за станом замовлення, узгодити нові параметри доставки, перерахувати орієнтовний час прибуття та надати відповідь, яка буде логічно пояснювати результат.

Традиційний шлюз API зазвичай розв'язує задачу статичної маршрутизації: він приймає запит і перенаправляє його на потрібну кінцеву точку за наперед заданими правилами (шлях, метод, заголовки), а також може виконувати контроль доступу, обмеження частоти,

протоколювання та базове перетворення повідомлень [2]. Однак такий підхід не має механізмів семантичного розуміння природномовних звернень користувача і, як наслідок, «розумна» логіка переноситься до клієнтського застосунку або реалізується як жорстко запрограмована послідовність дій. Це знижує гнучкість системи, ускладнює супровід, збільшує час реалізації нових сценаріїв і підвищує ризик помилок, коли бізнес-правила змінюються частіше, ніж оновлюється клієнтська частина.

У роботі розглянуто можливість підсилення шлюзу API можливостями великих мовних моделей, здатних виконувати семантичний аналіз тексту, виділяти наміри та сутності (наприклад, номер замовлення, прізвище отримувача, адресу, населений пункт, час), а також будувати план дій і взаємодіяти із зовнішніми функціями через керований виклик інструментів. У контексті трекінгу замовлень це означає, що користувач може формулювати запит природною мовою, а система має перетворити його на послідовність коректних дій над мікросервісами. Важливою перевагою є підтримка багатокрокової взаємодії: якщо у запиті бракує ідентифікатора замовлення або даних для виконання операції, модель формує уточнювальне запитання, після чого продовжує виконання, зберігаючи контекст сесії. Таким чином підвищується «людяність» взаємодії та зменшується кількість ситуацій, коли користувач отримує формальну відмову через неповні дані.

Метою роботи є підвищення гнучкості та ефективності взаємодії користувачів з інформаційною системою трекінгу замовлень шляхом розробки методу інтелектуальної маршрутизації, що дозволяє автоматично трансформувати запити, надані природною мовою, у послідовність викликів відповідних мікросервісів. Об'єктом досліджень є процеси маршрутизації та оркестрації запитів у розподілених інформаційних системах з мікросервісною архітектурою, а предметом – методи та програмні засоби інтелектуальної маршрутизації на основі семантичного аналізу запитів користувачів із застосуванням великих мовних моделей.

Запропонований метод ґрунтується на поданні кожного мікросервісу або його кінцевих точок як «інструментів» із формалізованим описом у вигляді метаданих [3]. У таких метаданих фіксуються призначення дії, перелік параметрів, їх типи, допустимі значення та обмеження, а також правила доступу й очікувана структура відповіді. На основі цього опису велика мовна модель виконує семантичний розбір запиту: визначає намір користувача, витягує потрібні сутності, співвідносить їх із доступними інструментами та формує план виконання у вигляді послідовності кроків. Принципово важливо, що модель не повинна безпосередньо звертатися до мережі або виконувати виклики мікросервісів напряду: вона лише пропонує структурований план і параметри, а фактичне виконання відбувається в контрольованому виконавчому шарі шлюзу. Саме такий

поділ відповідальностей дозволяє поєднати гнучкість семантичного аналізу з передбачуваністю та керованістю виконання.

Для забезпечення коректної роботи у реальному середовищі запропоновано виділяти в архітектурі інтелектуального шлюзу три узгоджені рівні: безпековий, інтелектуальний та виконавчо-інтеграційний. Безпековий рівень відповідає за автентифікацію, авторизацію, обмеження частоти запитів, захист від зловживань і первинне протоколювання. Інтелектуальний рівень відповідає за інтерпретацію природномовного звернення та побудову плану дій. Виконавчо-інтеграційний рівень забезпечує детерміноване виконання плану в мікросервісному середовищі, а також збір і нормалізацію відповідей. Таке структурування важливе з погляду супроводу: кожен рівень може розвиватися незалежно, має свої критерії якості та набір тестів, а ризики від помилок моделі мінімізуються за рахунок жорсткого контролю виконання.

Наукова новизна роботи полягає у розробці методу семантичної маршрутизації на основі великих мовних моделей, який, на відміну від традиційних підходів із жорсткою логікою перенаправлення, дозволяє динамічно визначати необхідний набір мікросервісів та параметрів їх виклику залежно від змісту та контексту запиту, забезпечуючи автоматичну оркестрацію складних сценаріїв взаємодії, що не були заздалегідь описані у статичній конфігурації шлюзу. Важливою відмінністю запропонованого підходу є акцент на безпечному виконанні: модель не є виконавцем, а лише «планувальником», тоді як контроль доступу, перевірка параметрів і політики виконання залишаються на інфраструктурному рівні.

У підсумку запропонований підхід перетворює шлюз API з інфраструктурного компонента статичної маршрутизації на інтелектуальний шар взаємодії, який здатний інтерпретувати запит користувача, формувати послідовність дій і виконувати її під повним контролем політик безпеки та експлуатаційних обмежень.

Список використаних джерел:

1. Christopher M. Logistics & Supply Chain Management. Pearson UK, 2016. 525 p.
2. Richardson C. Microservices Patterns: With Examples in Java. Shelter Island: Manning Publications, 2018. 642 p.
3. Iusztin P., Labonne M. LLM Engineer's Handbook: Master the art of engineering large language models from concept to production. Packt Publishing, 2024. 522p.