

УДК 004.8

Поляков О.М.

студ. гр. КНТ-214м НУ «Запорізька політехніка»

ЗАСТОСУВАННЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ ІНТЕЛЕКТУАЛЬНОЇ МАРШРУТИЗАЦІЇ ЗВЕРНЕНЬ У МЕСЕНДЖЕР-КАНАЛАХ

Проблема. Бізнеси, що працюють одразу в кількох месенджер-каналах, стикаються з однією і тією ж складністю — коректно розподілити потік вхідних повідомлень. Кнопкові меню клієнти ігнорують або заходять не в той розділ, а фільтри за ключовими словами ламаються при найменшій варіативності формулювань. За модельними розрахунками, близько третини

звернень потрапляє не туди – і це прямо впливає на завантаження операторів та загальне враження від сервісу.

Мета. Запропонувати підхід до автоматичної маршрутизації вхідних повідомлень, який працює з довільним текстом і не вимагає жорстких правил чи ручного налаштування сценаріїв.

Метод. В основі рішення – проміжний шар (middleware), реалізований на Python з використанням FastAPI. Перш ніж повідомлення дійде до оператора чи бота, його аналізує мовна модель. Запит іде через звичайний HTTP до хмарного inference endpoint, відповідь – строго структурований JSON: intent, entities, confidence. Жодного вільного тексту модель не генерує – замість цього вона через механізм tool use безпосередньо запускає потрібну бізнес-логіку системи: звертається до CRM, перевіряє розклад, відкриває тикет. На практиці одне повідомлення обробляється за 280–420 мс – цього цілком достатньо для живого діалогу в месенджері. Навіть при 50 паралельних запитах затримка зростає не більш ніж на 15%. Список intent-класів (до 20 штук) змінюється через системний промпт, без жодного перенавчання моделі.

Результат. Система вже працює в реальних умовах – підключено понад 5 CRM-платформ різного профілю (облік клієнтів, онлайн-запис, helpdesk), загальна кількість підключених компаній перевищила 1000. Модельні розрахунки для сервісного бізнесу показують: частка звернень, вирішених без оператора, зростає з 20% до 65%, час першої відповіді падає з 5 хвилин до секунди, а помилки маршрутизації – з 35% до 4%. Вартість обробки одного звернення через API – менше \$0.001. Канали і intent-класи додаються без змін у логіці системи.