

Проектування та створення веб-застосунку на основі поєднання можливостей **LLM** та **RAG**

Виконав студент: Українець Дмитро Сергійович

Науковий керівник: Олецький Олексій Віталійович

Актуальність та мета роботи

Актуальність: широкого поширення набуло використання LLM у задачах генерації тексту. Водночас такі моделі можуть формувати надто загальні відповіді або недостатньо враховувати предметний контекст. RAG дає змогу доповнювати генерацію релевантною інформацією.

Мета: проєктування та створення веб-застосунку на основі поєднання можливостей LLM та RAG, який на основі результатів аудиту формує структуровані рекомендації щодо усунення виявлених проблем.

Огляд ключових технологій

RAG (Retrieval-Augmented Generation) - підхід, що поєднує пошук релевантного контексту у зовнішніх джерелах інформації та подальшу генерацію відповіді. Він допомагає зробити результат точнішим і більш прив'язаним до предметної області.

LLM (Large Language Model) - велика мовна модель, навчена на значних обсягах текстових даних. Вона аналізує вхідний текст, враховує контекст запиту та формує зв'язну відповідь природною мовою.

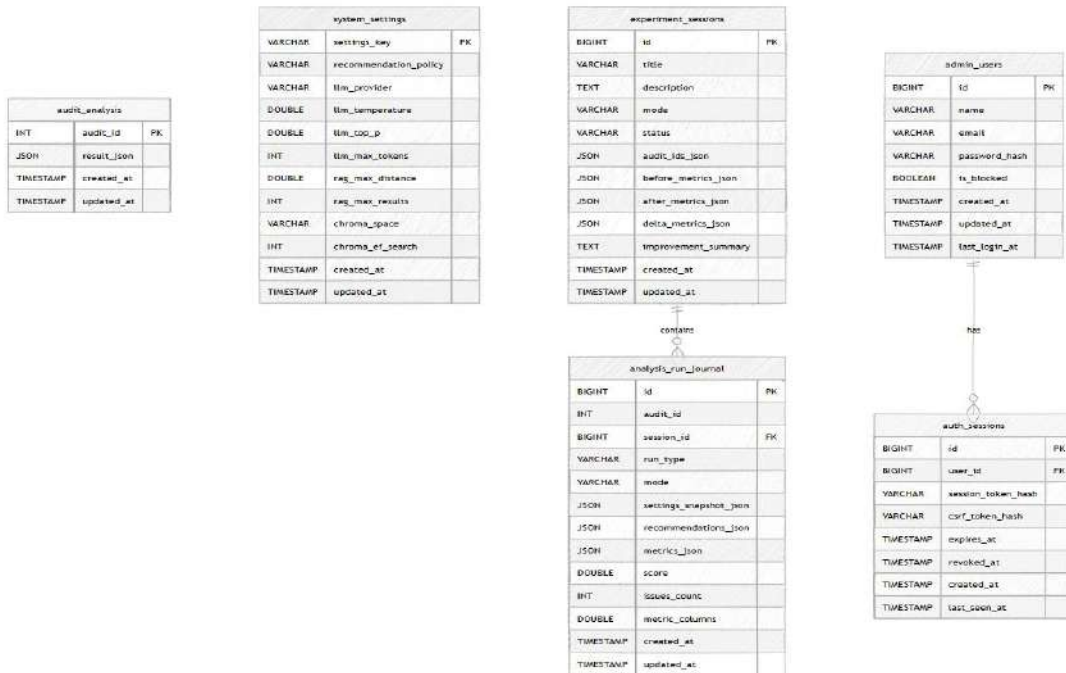
Загальна схема роботи **RAG**-системи



Стек технологій

Клієнтська частина	Серверна частина	LLM	DB
• JavaScript • Vue 3 • Vite • Tailwind CSS	• Python • FastAPI	• OpenAI • Gemini	• MySQL • ChromaDB

Структура бази даних



Реалізація серверної частини

01

Отримання
результатів

02

Нормалізація
даних

03

Виділення
проблемних пунктів

04

Пошук релевантного
контексту

05

Формування
рекомендацій

06

Збереження результатів
в БД і кеші

РЕАЛІЗАЦІЯ КЛІЄНТСЬКОЇ ЧАСТИНИ

Публічна частина: головна сторінка



Головна

Введіть ідентифікатор аудиту, щоб отримати рекомендації.

Ідентифікатор аудиту

Наприклад: 12345

Проаналізувати

Формування рекомендацій може зайняти 1-2 хвилини залежно від навантаження.

Приклад бізнес-
профілю на Google
Картах

РЕАЛІЗАЦІЯ КЛІЄНТСЬКОЇ ЧАСТИНИ

Публічна частина: сторінка рекомендацій

[← На головну](#)[Вхід](#)[Довідка](#)

Рекомендації для профілю на Google Картах

Результати аналізу аудиту #89

[Оновити рекомендації](#)

64

Загальний бал

Добре

Виявлено проблем

18

Проблеми та рекомендації

ПОШУК

КАТЕГОРІЯ

Усі категорії

ДЖЕРЕЛО

Усі джерела

ПРІОРИТЕТ

Усі

ВПЛИВ

Усі

СКЛАДНІСТЬ

Усі

МІНІМАЛЬНА ВАГА

Наприклад: 10

Допустимий діапазон: 1–100.

СОРТУВАТИ ЗА

Вага

НАПРЯМ СОРТУВАННЯ

За спаданням

Знайдено: 18

[Скинути фільтри](#)

Проблема

Вага: 3

У картці компанії є активна кнопка "Замовити" та/або "Забронювати онлайн зустріч" та/або посилання на альтернативні методи зв'язку?

Ви не використовуєте можливість кнопки "Замовити" чи "Забронювати", яка дозволяє клієнтам миттєво зв'язатися з вами. Рекомендуємо додати посилання на месенджери, соціальні мережі або сервіс онлайн-бронювань через сайт.

Рекомендація

Джерело: База правил

Низький пріоритет

Вплив ●●●

Складність: висока

Проблема виявлена під час автоматичного аудиту бізнес-профілю.
Рекомендована дія сформована на основі бази правил.

Що зробити: Перевірити, чи додано у профіль компанії кнопку "Замовити", "Забронювати", "Записатися" або інше інтерактивне посилання для здійснення дії користувачем.
Це може бути посилання на сервіс онлайн-бронювання, форму запису на сайті, або навіть інтегроване рішення в Google Business Profile.

Категорія: активіти

РЕАЛІЗАЦІЯ КЛІЄНТСЬКОЇ ЧАСТИНИ

Публічна частина: довідка та авторизація

Довідкова інформація

Пояснення полів, джерел рекомендацій та принципів формування.

Загальні показники

- **Загальний бал** — підсумкова оцінка профілю за результатами перевірок у діапазоні від 0 до 100. Чим більше суттєвих проблем виявлено, тим нижчий бал.
- **Кількість проблем** — кількість перевірок, у яких система виявила недоліки або зауваження.

Блок «Проблема»

- **Назва проблеми** — коротко описує, яку саме невідповідність виявлено в профілі.
- **Опис проблеми** — пояснює, що саме не так і чому це може впливати на якість профілю або його видимість.
- **Вага** — показує відносну важливість проблеми. Чим більше значення, тим більше уваги варто приділити її виправленню.

Вага проблеми формується на основі результатів окремої перевірки та використовується для сортування і пріоритезації рекомендацій.

Блок «Рекомендація»

- **Пояснення** — коротко описує, чому система пропонує саме таке виправлення.
- **Джерело** — показує, звідки взята фінальна рекомендація: з бази правил, від LLM або з резервного механізму.

×

Увійти в систему

×

Пошта

Формат пошти: name@example.com

Пароль

Щонайменше 8 символів, серед яких хоча б 1 літера і 1 цифра.

Скасувати

Увійти

Адмін-частина: сторінка налаштувань

Системні налаштування

×

Режим роботи системи

☐ База правил

☐ RAG

☒ LLM

Провайдер LLM

☒ OpenAI

☐ Gemini

LLM_TEMPERATURE

0.25

Менше значення дає стабільніші та передбачуваніші відповіді.

LLM_TOP_P

0.95

Обчислює вибір моделі найбільш імовірними токенами.

LLM_MAX_TOKENS

720

Діапазон: 1 – 4000.

RAG_MAX_DISTANCE

0.18

Для ір значення підбирається експериментально. Використовуйте невеликі порогові та перевіряйте якість результатів.

RAG_MAX_RESULTS

5

Діапазон: 1 – 5.

CHROMA_SPACE

i

ір

▼

Простір відстані для векторного пошуку.

CHROMA_EF_SEARCH

i

321

Діапазон: 1 – 1000.

Зберегти

Зберегти та оновити

Адмін-частина: історія запусків аудиту

Історія запусків аудиту

Перегляд попередніх запусків, налаштувань, рекомендацій і метрик.

Запуски

#9

rule_base

Тип запуску: experiment
2026-04-12 21:42:28

#4

rule_base

Тип запуску: experiment
2026-04-12 21:38:00

#3

rag

Тип запуску: interactive
2026-04-12 18:09:49

#2

llm

Тип запуску: interactive
2026-04-12 18:07:22

#1

rule_base

Тип запуску: interactive
2026-04-12 18:04:18

РЕЖИМ

rule_base

ТИП ЗАПУСКУ

experiment

ДАТА

2026-04-12 21:42:28

Ключові метрики

Підсумковий optimization score

—

Частка аудитів із RAG-кандидатами

—

Покриття обґрунтованими RAG-
відповідями

—

Частка fallback-відповідей

—

Середній обернений ранг точного
правила

—

Доступність точного правила

—

Частка використання точного правила

—

Середня релевантність правила

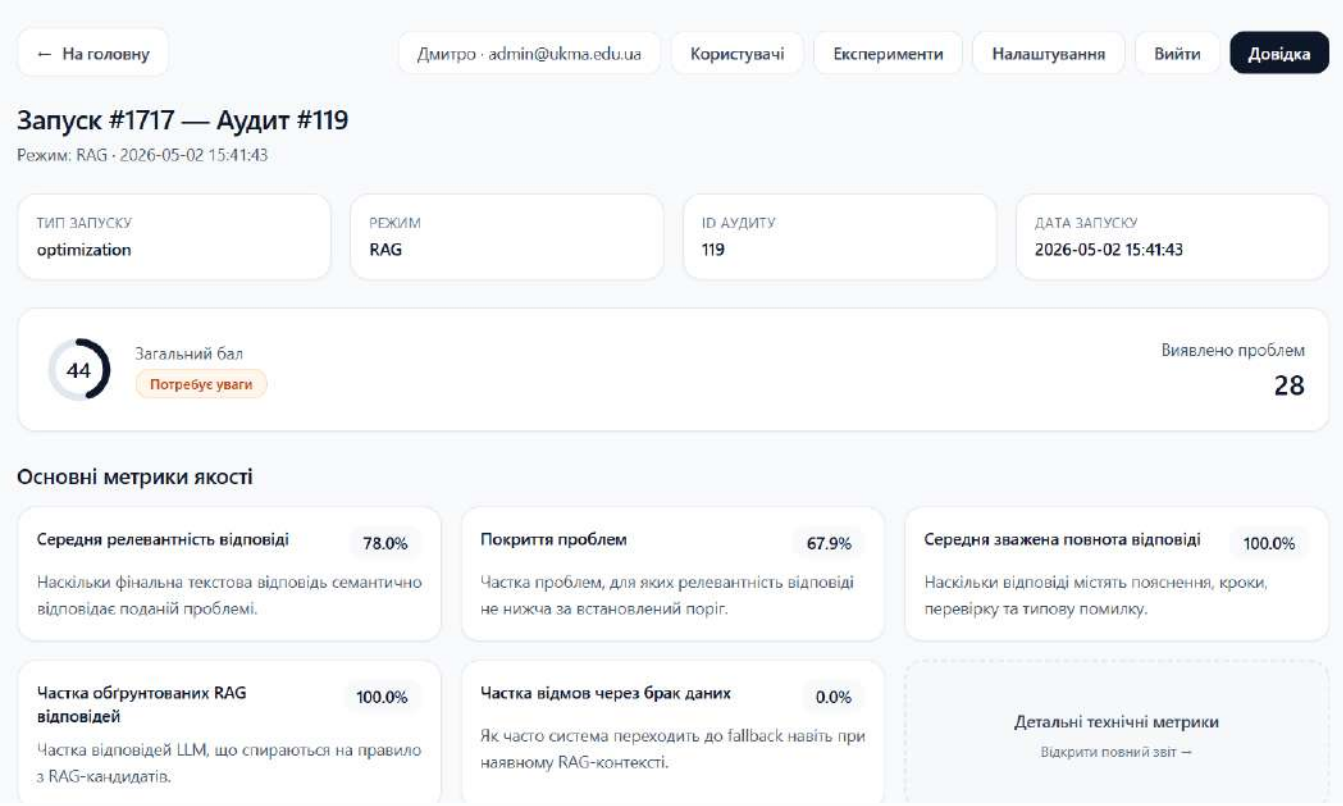
79.3%

Середня зважена повнота правила

81.1%

Snapshot налаштувань

Адмін-частина: сторінка детального аналізу



Адмін-частина: експериментальний модуль

[← На головну](#)

Дмитро · admin@ukma.edu.ua

Користувачі

Експерименти

Налаштування

Вийти

Довідка

Експериментальні сесії

Створення, запуск і перегляд дослідницьких серій аналізу аудитів.

Нова експериментальна сесія

Режим експериментальної сесії автоматично застосовує поточні налаштування системи.

Назва сесії

Наприклад: Підбір параметрів режиму RAG

Опис

Коротко опишіть мету або умови цього експерименту

Ідентифікатори аудитів

Наприклад: 89, 178, 245 (через кому, пробіл або новий рядок)

ПОТОЧНИЙ РЕЖИМ

LLM

Створити сесію

Адмін-частина: керування користувачами

Управління користувачами

Створити нового користувача

Ім'я

Ім'я користувача

Пошта

name@example.com

Формат пошти: name@example.com

Пароль

Новий пароль

Щонайменше 8 символів, серед яких хоча б 1 літера і 1 цифра.

Показати пароль

Створити користувача

Адмін

a.admin@ukma.edu.ua

Активний

ID: 2

Заблокувати

Профіль

Ім'я

Адмін

Пошта

a.admin@ukma.edu.ua

Зміна пароля

Новий пароль

Новий пароль

Щонайменше 8 символів, серед яких хоча б 1 літера і 1 цифра.

Показати пароль

Дмитро

admin@ukma.edu.ua

Активний

ID: 1

Поточний користувач

Заблокувати

Профіль

Ім'я

Дмитро

Пошта

admin@ukma.edu.ua

Зміна пароля

Новий пароль

Новий пароль

Щонайменше 8 символів, серед яких хоча б 1 літера і 1 цифра.

Показати пароль

Режими роботи системи

База правил	RAG	LLM
Модель векторного подання тексту: - all-MiniLM-L6-v2 Налаштування семантичного пошуку: - Векторний простір подібності (cosine/l2/ip); - К-сть результатів (1/2/3 ...); - Порогова дистанція релевантності (наприклад, 0.35 для cosine, залежить від векторного простору подібності); - К-сть сусідів під час пошуку (50/100/200 ...).	База правил	Провайдери LLM: - OpenAI GPT-4o mini - Google Gemini 2.5 Flash-Light
	+	
	LLM-складова	

Експериментальне дослідження

Мета: оцінити якість рекомендацій залежно від режиму роботи системи.

Вхідні дані: 20 аудитів.

Автоматичний підбір параметрів за допомогою Optuna:

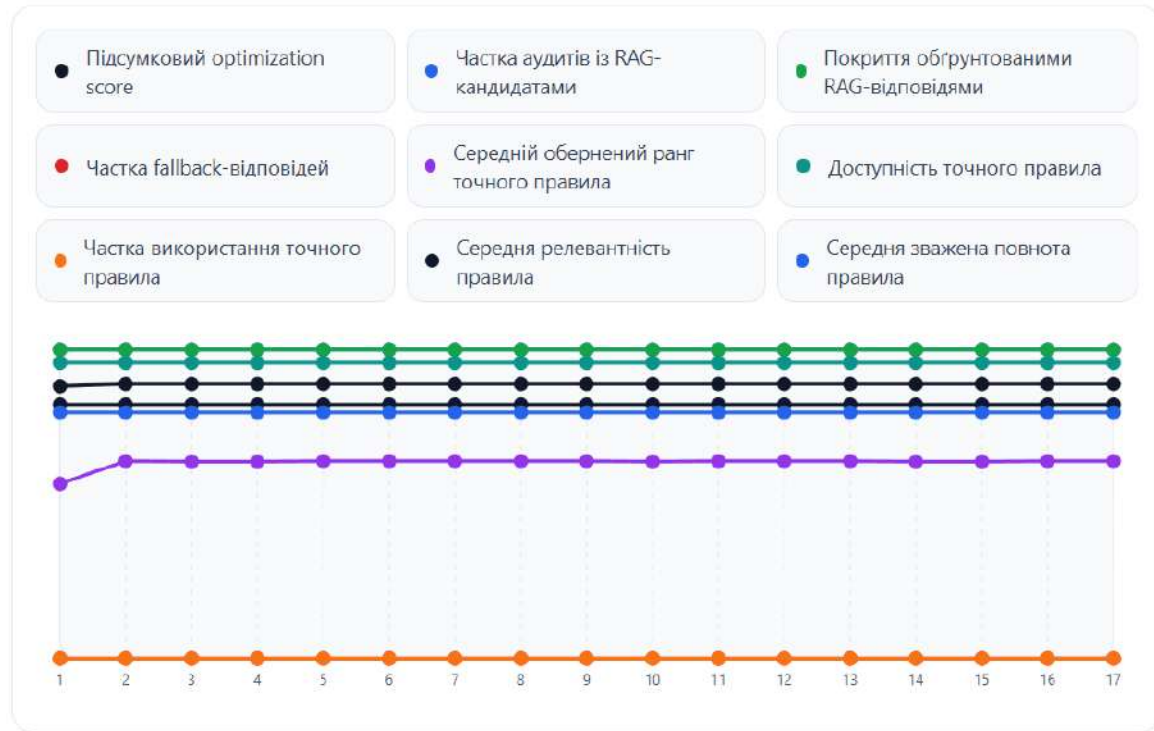
- TPESampler - добір нових конфігурацій параметрів на основі попередніх результатів;
- HyperbandPruner - дострокове зупинення неперспективних запусків.

Критерій оцінювання: найбільший підсумковий оптимізаційний бал.

Порівнювані режими: База правил / LLM / RAG.

Результати дослідження режиму бази правил

Динаміка метрик за всіма запусками



Вхідні дані:

- 17 завершених запусків
- 13 достроково зупинених запусків

Найкращий бал: 88,89 %

Висновок: режим працює стабільно, але якість результатів залежить від повноти, уніфікованості та деталізації правил.

Результати дослідження режиму LLM

Динаміка метрик за всіма запусками



Вхідні дані:

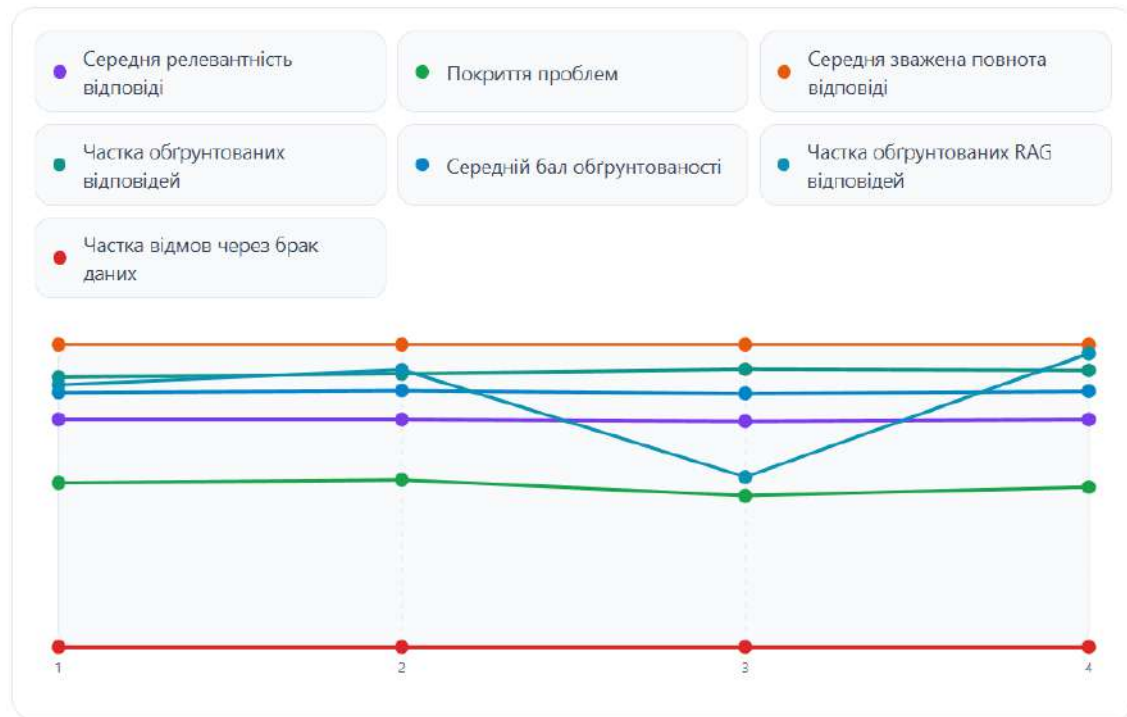
- 4 завершені запуски
- 11 достроково зупинених запусків

Найкращий бал: 85,90 %

Висновок: режим LLM працює технічно стабільно та формує повні відповіді, але є менш контрольованим, оскільки частина відповідей не досягає потрібного рівня релевантності до виявленої проблеми.

Результати дослідження режиму **RAG**

Динаміка метрик за всіма запусками



Вхідні дані:

- 4 завершені запуски
- 10 достроково зупинених запусків
- 1 запуск завершився помилкою

Найкращий бал: 86,06 %

Висновок: режим RAG

забезпечує високу обґрунтованість відповідей і повну структурну наповненість рекомендацій, але результат значною мірою залежить від якості добору релевантного контексту.

LLM та RAG: переваги та недоліки

Метрика	LLM	RAG
Частка успішних відповідей LLM	98,43 %	98,39 %
Частка фінальних LLM-відповідей	98,43 %	98,39 %
Частка резервних відповідей після LLM	1,57 %	1,61 %
Середня релевантність відповіді	76,25 %	75,28 %
Покриття проблем	54,92 %	52,91 %
Частка обґрунтованих відповідей	90,74 %	91,50 %
Середній бал обґрунтованості	83,02 %	84,62 %
Підсумковий оптимізаційний бал	85,90 %	86,06 %

Підсумковий бал:

- RAG кращий на **0,16 %**

Переваги режиму LLM:

- частка успішних відповідей LLM: **+0,04 %**
- частка фінальних LLM-відповідей: **+0,43 %**
- частка резервних відповідей після LLM: **-0,04 %**
- середня релевантність відповіді: **+0,97 %**
- покриття проблем: **+2,01 %**

Переваги режиму RAG:

- частка обґрунтованих відповідей: **+0,76 %**
- середній бал обґрунтованості: **+1,60 %**

Висновок: режим LLM має дещо більше кількісних переваг, але частина з них є дуже незначними, хоча LLM формує релевантніші до тексту проблеми відповіді; однак RAG забезпечує кращу обґрунтованість результату та має вищий підсумковий оптимізаційний бал, тому є доцільнішим як основний режим формування рекомендацій.

Порівняння відповіді RAG та LLM

Проблема Важк: 16

У компанії є товари у профілі компанії?

У вашому профілі відсутні товари. Через це компанія не бере участі в пошуковій системі Google Товари і втрачає можливість показувати свої позиції потенційним клієнтам. УВАГА! Не є важливим, якщо компанія додала товари у Merchant Center!

Рекомендація

Джерело: модель Низький пріоритет Вплив ●●● Складність: низька

У профілі компанії відсутні товари, що заважає компанії брати участь у пошуковій системі Google Товари та показувати свої позиції потенційним клієнтам.

Що зробити:

1. Додайте 5–10 товарів/позицій з фото, назвою, коротким описом і ціною/діапазоном.
2. Для кожного товару додайте посилання на відповідну сторінку сайту (якщо є) або на сторінку з деталями.
3. Оновлюйте інформацію про товари після змін.

Перевірка: Переконайтеся, що товари відображаються у профілі компанії в Google.

Типова помилка: Ігнорування додавання товарів або недостатня кількість інформації про них.

Категорія: general

Проблема Важк: 16

У компанії є товари у профілі компанії?

У вашому профілі відсутні товари. Через це компанія не бере участі в пошуковій системі Google Товари і втрачає можливість показувати свої позиції потенційним клієнтам. УВАГА! Не є важливим, якщо компанія додала товари у Merchant Center!

Рекомендація

Джерело: модель Високий пріоритет Вплив ●●● Складність: середня

У вашому профілі відсутні товари, що заважає компанії брати участь у пошуковій системі Google Товари. Це важливо для підвищення видимості та залучення потенційних клієнтів.

Що зробити:

1. Додайте товари у профіль компанії.
2. Переконайтеся, що товари мають коректні описи та зображення.
3. Оновіть інформацію про ціни та наявність товарів.

Перевірка: Переконайтеся, що товари відображаються у профілі компанії в Google.

Типова помилка: Не додавати достатню інформацію про товари або ігнорувати їх оновлення.

Категорія: content

```
{  
  "id": "Оптимізація профілю  
компанії в Google:has_products",  
  "title": "У компанії є товари у  
профілі компанії?",  
  "description": "Розділ «Товари»  
дозволяє показати асортимент прямо в  
профілі та вести трафік на сайт.",  
  "recommendation": "Додайте 5–10  
товарів/позицій з фото, назвою,  
коротким описом і ціною/діапазоном.  
Для кожного товару додайте посилання  
на відповідну сторінку сайту (якщо  
є) або на сторінку з деталями.",  
  "category": "general",  
  "priority": "low",  
  "impact": "medium",  
  "difficulty": "low"  
}
```

Перспективи розвитку



Інтеграція та експлуатаційна готовність

- Інтеграція з Mappers GEO Tools
- Підготовка до роботи в реальному середовищі
- Посилення надійності та безпеки системи



Покращення якості рекомендацій

- Розширення та уніфікація бази правил
- Удосконалення добору релевантного контексту
- Тестування моделей, параметрів і конфігурацій RAG та LLM



Розвиток клієнтської частини

- Порівняння результатів різних режимів
- Експорт результатів аналізу
- Оцінювання рекомендацій користувачем

Висновки

Реалізований веб-застосунок:

- обробляє готові результати аудиту бізнес-профілю на Google Картах;
- виділяє виявлені проблеми та добирає релевантні правила;
- формує структуровані рекомендації в різних режимах роботи.

Результати експериментів:

- LLM-режим - краща релевантність та покриття проблем у відповідях;
- RAG-режим - краща обґрунтованість і зв'язок із підготовленими правилами.

Практичне значення: розроблений веб-застосунок демонструє, що поєднання RAG та LLM дає змогу посилити зв'язок сформованих відповідей із підготовленими правилами та предметним контекстом, що підвищує їхню обґрунтованість і керованість.

Дякую за увагу!

Буду радий відповісти на ваші запитання.

d.ukrainets@ukma.edu.ua

CREDITS: This presentation template was created by Slidesgo, and includes icons, infographics & images by Freepik