

ДОСЛІДЖЕННЯ МЕТОДІВ ВІЗУАЛІЗАЦІЇ ТА АНАЛІЗУ ІНФОРМАЦІЇ З ВИКОРИСТАННЯМ LLM

Будник М.О. ІПЗм-23-3
Керівник: Доц. Голян В.В.

АКТУАЛЬНІСТЬ ТЕМИ ДОСЛІДЖЕННЯ

- Експоненційне зростання обсягів даних вимагає ефективних інструментів для їх аналізу та візуалізації.
- Існуючі рішення (Tableau, Power BI) часто орієнтовані на технічних користувачів, мають складний інтерфейс або обмежену підтримку автоматизованого аналізу.
- Це обмежує доступність аналітики для широкого кола користувачів, особливо без навичок програмування.
- Великі мовні моделі (LLM) пропонують потенціал для автоматизації цих процесів, але їх інтеграція в аналітичні системи залишається недостатньо дослідженою.
- Автоматизація аналізу за допомогою LLM може спростити отримання інсайтів та зробити аналітику доступною для некваліфікованих користувачів.

МЕТА ТА ЗАДАЧІ ДОСЛІДЖЕННЯ

- **Мета роботи:** Дослідження принципів застосування LLM для аналізу та візуалізації даних, розробка концептуальної архітектури та програмного прототипу такої системи, дослідження методів генерації інсайтів та візуалізацій, порівняння ефективності різних моделей LLM (зокрема GPT-4o, Claude 3, Gemini Pro) та дослідження оптимальних технік побудови пром프트ів для даної задачі.
- **Задачі дослідження:**
 - Проаналізувати існуючі підходи до візуалізації та аналізу даних.
 - Розробити архітектуру прототипу системи для аналізу даних з різних джерел (CSV, JSON) та автоматизованого створення візуалізацій за допомогою LLM.
 - Дослідити методи аналізу та візуалізації інформації за допомогою LLM.
 - Описати алгоритми аналізу та візуалізації даних з використанням LLM.
 - Дослідити проблеми використання LLM для аналізу даних (конфіденційність, достовірність, обмеження).
 - Провести порівняльний аналіз використання різних LLM.

ОБ'ЄКТ ТА ПРЕДМЕТ ДОСЛІДЖЕННЯ

- **Об'єкт дослідження:** Процес дослідження та розробки системи для візуалізації та аналізу даних з використанням великих мовних моделей (LLM). (Альтернативно з ВСТУПУ: Процес обробки, аналізу та візуалізації даних з використанням великих мовних моделей)
- **Предмет дослідження:** Методи візуалізації та аналізу інформації, що використовують LLM для автоматизації процесу пошуку інсайтів та візуалізації, а також теоретичні основи її функціонування.

НАУКОВА НОВИЗНА

- Розроблено методику генерації інтерактивних візуалізацій на основі динамічних промптів, адаптованих під специфіку даних.
- Встановлено залежність між якістю промптів та точністю результатів аналізу.

ОГЛЯД ІСНУЮЧИХ РІШЕНЬ ТА ЛІТЕРАТУРИ

- **Існуючі інструменти:**
 - Datadog, Grafana, Google Data Studio, Microsoft Power BI, Tableau.
 - Переваги: Потужні можливості візуалізації, інтеграції.
 - Недоліки: Орієнтація на технічних користувачів, складність налаштування, обмежений автоматизований аналіз, вартість.
- **Літературний огляд:**
 - Принципи роботи LLM (Transformer, attention mechanisms), їх застосування в аналізі даних.
 - Порівняння LLM: Gemini (мультимодальність), GPT (обробка тексту), Llama (відкритість, ефективність).
 - Візуалізація даних на JavaScript (D3.js, Chart.js, Plotly.js).
 - Методи аналізу даних з LLM (NLP, ML).
 - Актуальність підтверджена аналізом джерел, виявлено потребу в дослідженні інтеграції LLM та prompt engineering для аналітичних систем.

ПОСТАНОВКА НАУКОВО-ТЕХНІЧНОЇ ЗАДАЧІ

Розробити інструмент автоматизованого аналізу даних із застосуванням LLM, що забезпечує:

- Підтримку структурованих (CSV, JSON) і текстових форматів даних.
- Генерацію інсайтів і візуалізацій за допомогою природної мови.
- Вбудовану перевірку достовірності результатів (зважаючи на "галюцинації" LLM).
- Порівняльний аналіз декількох LLM (GPT-4, Claude 3, Gemini Pro) з використанням об'єктивних метрик точності, вартості та часу відгуку.

Обмеження дослідження:

- Аналіз 3-х LLM (GPT-4o, Claude 3 Opus, Gemini 1.5 Pro).
- Тестування на публічних датасетах.
- Фокус на текстових та структурованих даних.

ТЕОРЕТИЧНЕ ДОСЛІДЖЕННЯ: LLM В АНАЛІЗІ ДАНИХ

Архітектурні особливості LLM (GPT-4, Claude 3, Llama 3):

- Базуються на архітектурі трансформерів з механізмами уваги.
- Дозволяють аналізувати зв'язки в структурованих даних, генерувати семантичні уявлення для неструктурованих текстів, адаптуватися до різних мовних шаблонів.

Обробка структурованих даних (CSV, JSON):

- Дані серіалізуються в текстовий формат.
- Важливість формату вводу та "structured prompting" для точності.
- Використання режимів виводу JSON або "function calling" для гарантування коректності структури.

Аналіз текстових даних:

- Класифікація тексту, аналіз настрою (few-shot, zero-shot).
- Розпізнавання іменованих сутностей (NER) – складніше для загальних LLM порівняно зі спеціалізованими моделями.

ПОРІВНЯННЯ LLM ТА РОЛЬ ПРОМПТ-ІНЖИНІРИНГУ

Порівняльні характеристики LLM:

- GPT-4o (OpenAI): Економічність, високі оцінки на бенчмарках MMLU, MGSM, HumanEval.
- Claude 3 (Anthropic): Надійність, мультидисциплінарність, обробка візуальної інформації (Opus, Sonnet, Haiku).
- Gemini (Google): Масштабованість, підтримка довгого контексту (до 1M токенів у Gemini 1.5 Pro).
- Вибір залежить від балансу "швидкість-якість-ціна" та специфіки завдання.

Роль prompt engineering:

- Ключовий для підвищення продуктивності LLM.
- Техніки: Zero-shot, Few-shot, Chain-of-Thought (CoT), system prompts.
- CoT значно покращує результати у складних задачах.
- Правильна побудова промπτу критична для точності та уникнення "галюцинацій".

Обмеження LLM та проблема "галюцинацій":

- Обмежений розмір контекстного вікна.
- Негарантована достовірність відповідей ("галюцинації").
- Вартість та ресурсоемність.
- Стратегії боротьби: Retrieval-Augmented Generation (RAG), самодіагностика, заохочення чесності.

РОЗРОБКА ПРОТОТИПУ: АРХІТЕКТУРА ТА ТЕХНОЛОГІЇ

Розроблено веб-додаток для аналізу даних з використанням LLM.

Архітектура: Модульна, забезпечує гнучкість інтеграції різних LLM.

Технологічний стек:

- Next.js: Фреймворк для фронтенду та бекенду (серверний рендеринг, API-маршрути).
- AI SDK (Vercel AI SDK): Уніфікована взаємодія з LLM (OpenAI, Anthropic, Google).
- ShadcnUI: Створення користувацького інтерфейсу.
- Recharts: Побудова інтерактивних графіків.
- Zod: Валідація схем даних та типобезпека.

ФУНКЦІОНАЛЬНІ МОЖЛИВОСТІ

11

Домашня сторінка: Відображення карток раніше створених візуалізацій.

Створення нової візуалізації:

- Завантаження файлу даних (JSON, CSV, до 10 МБ).
- Вибір LLM (GPT-4o, Claude 3 Opus, Gemini 1.5 Pro).
- Вибір мови результатів (українська, англійська).
- Автоматичне скорочення даних при перевищенні лімітів токенів моделі.

Сторінка візуалізації та аналізу даних:

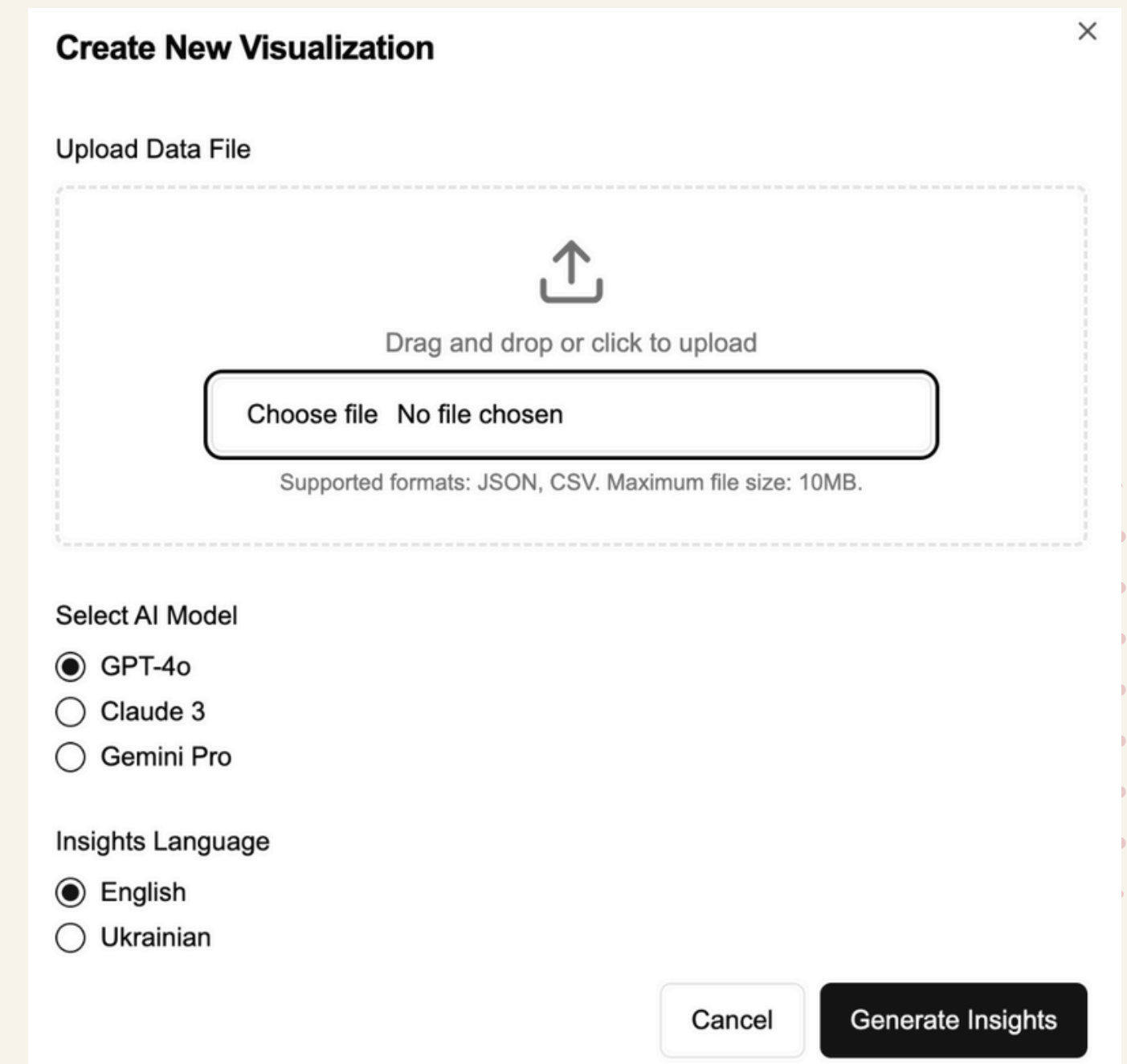
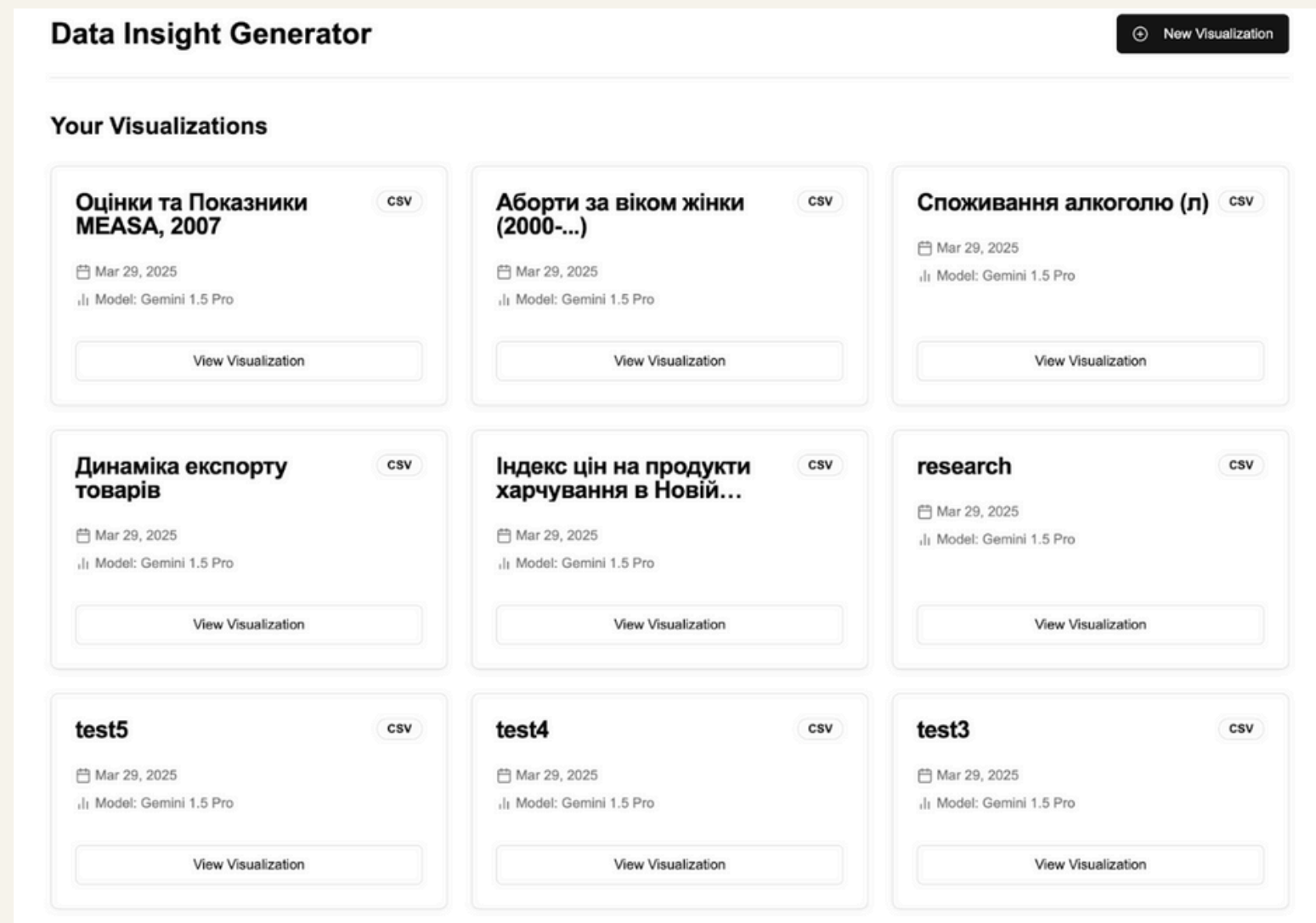
- Динамічно згенеровані віджети: графіки (різні типи), таблиці, текстові інсайти, метрики, загальний висновок.
- Кожен віджет супроводжується заголовком, описом та текстовими інсайтами.
- Можливість перегляду та завантаження вихідних даних.

Взаємодія з LLM API:

- Обробка та підготовка даних (csv-parse, truncateData).
- Формування промптів (заголовковий, для аналізу даних).
- Використання Zod для валідації відповідей LLM.
- AI SDK для уніфікованої взаємодії з OpenAI, Anthropic, Google.
- Метод generateObject для отримання структурованих відповідей.
- Динамічна трансформація даних для віджетів (JavaScript-функції dataTransform на клієнті).

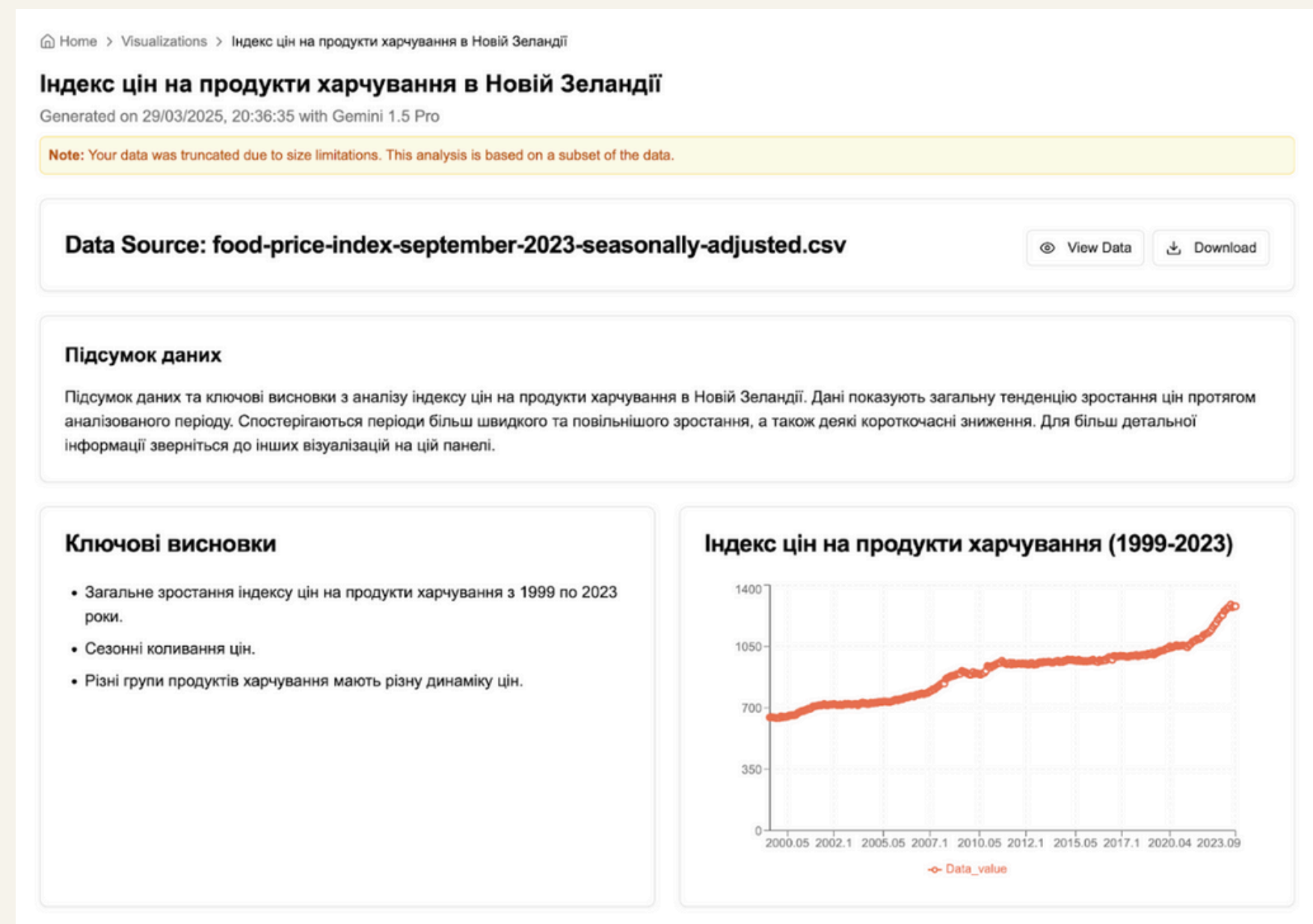
ДЕМОНСТРАЦІЯ ПРОТОТИПУ

- **Домашня сторінка:** Огляд створених візуалізацій, швидкий доступ до результатів.
- **Діалогове вікно створення візуалізації:** Завантаження даних, вибір моделі LLM та мови.



СТОРІНКА ВІЗУАЛІЗАЦІЇ

- Відображення результатів аналізу у вигляді інтерактивних віджетів: графіки, таблиці, інсайти.
- Інформаційна панель: назва, дата, модель, індикатор скорочення даних.
- Приклад віджетів: підсумок даних, ключові висновки, графік (напр., індекс цін).



ПОБУДОВА ПРОМПТУ ДЛЯ LLM

Розроблено ефективний промпт для керування LLM у процесі аналізу та візуалізації даних. Ключові компоненти промπτу:

1. Встановлення ролі та контексту: "Ви експерт з візуалізації даних".
2. Надання вхідних даних: Дані у форматі JSON/CSV, повідомлення про скорочення (якщо є).
3. Чіткі інструкції: Проаналізувати дані, створити дашборд, повернути JSON-конфігурації віджетів, забезпечити функцію dataTransform, використовувати вказану мову.
4. Детальний опис шаблонів віджетів: Для діаграм, таблиць, інсайтів, метрик, зведення, з описом поля dataTransform.
5. Важливі рекомендації: Щодо валідності JavaScript, форматування даних, змістовних інсайтів, повернення лише валідного JSON.
6. Чіткий формат відповіді: Приклад очікуваного масиву JSON-об'єктів.

Функціональні можливості промπτу: Комплексний аналіз, підтримка різних типів візуалізацій, програмний компонент для трансформації даних.

РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТАЛЬНОГО ДОСЛІДЖЕННЯ

- Прототип тестовано на різних наборах даних (продажі, метрики активності, часові ряди, категоріальні дані).

Порівняння LLM:

- Claude 3 Opus: Найкращі результати в генерації змістовних текстових інсайтів. Найкращий вибір доречних типів візуалізацій.
- GPT-4o: Висока точність у математичних обчисленнях. Генерація коректних функцій трансформації даних (dataTransform).
- Gemini 1.5 Pro: Переваги у швидкості обробки великих наборів даних. Іноді генерує менш структуровані текстові описи.

Ключове спостереження:

- Точність і корисність аналізу значно залежать від якості промптів та структури вихідних даних. Прототип успішно генерує візуалізації та інсайти.

ВИСНОВКИ

- Проведено аналіз предметної галузі, виявлено актуальність автоматизації аналізу даних за допомогою LLM.
- Систематизовано теоретичні основи використання LLM (Gemini, GPT, Llama), методів візуалізації (D3.js, Recharts) та промпт-інженерії.
- Розроблено концептуальну архітектуру та програмний прототип системи на базі Next.js, AI SDK, Zod, Recharts. Прототип реалізує завантаження даних (CSV, JSON), автоматичну генерацію візуалізацій та інсайтів.
- Експериментально досліджено ефективність LLM: Claude 3 Opus (інсайти, типи візуалізацій), GPT-4o (математика, трансформації), Gemini 1.5 Pro (швидкість).
- Обґрунтовано ефективні методи взаємодії з LLM, розроблено деталізований промпт.
- Підтверджено гіпотезу про ефективність використання LLM для спрощення та автоматизації аналізу даних, особливо для користувачів без технічних навичок.

ПРАКТИЧНЕ ЗНАЧЕННЯ ТА НАПРЯМКИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Практичне значення:


- Створений прототип демонструє реальну реалізацію запропонованих підходів.
- Система може спростити аналіз даних для користувачів без спеціалізованих навичок.
- Запропоновані методи інтеграції LLM та побудови промптів можуть бути застосовані для створення нових інструментів.
- Може значно скоротити час на аналіз даних та підвищити якість рішень.

Напрямки подальших досліджень:

- Розширення типів візуалізацій та можливостей їх кастомізації.
- Обробка потокових даних в режимі реального часу.
- Дослідження складніших методів ML для аналізу залежностей у даних з інтеграцією LLM.
- Оптимізація алгоритмів для роботи з великими обсягами даних, масштабування системи.
- Розробка спеціалізованих промптів для конкретних типів аналізу та галузей.
- Покращення механізмів боротьби з "галюцинаціями" LLM та верифікації результатів.

АПРОБАЦІЯ РЕЗУЛЬТАТІВ

- Основні положення та висновки роботи представлено у статті «Дослідження методів візуалізації та аналізу інформації з використанням LLM».
- Стаття прийнята до публікації в №21 студентського наукового журналу «UNIVERSUM».
- Випуск заплановано на 20 червня 2025 року.

**МОЛОДІЖНА
НАУКОВА
ЛІГА** 

Громадська організація «Молодіжна наукова ліга».
Номер запису в Реєстрі громадських об'єднань: 1506433.
Адреса: вул. Зодчих, буд. 40, офіс 103; м. Вінниця, Вінницька обл., 21037
Організація функціонує як відокремлений підрозділ ТОВ «UKRLOGOS Group».
ЄДРПОУ: 44574526
IBAN: UA433052990000026002046104529
Банк ВФ АТ КБ «ПриватБанк»; МФО 44574526
Свідоцтво суб'єкта видавничої справи: ДК № 7860 від 22.06.2023.

Д О В І Д К А
ПРО ПРИЙНЯТТЯ СТАТТІ ДО ПУБЛІКАЦІЇ

22.05.2025

Шановний(і) автор(и):
Будник Максим Олексійович,

Редакційний комітет з радістю повідомляє, що стаття «Дослідження методів візуалізації та аналізу інформації з використанням LLM» прийнята до публікації в № 21 студентського наукового журналу «UNIVERSUM», випуск якого заплановано на 20 червня 2025 року.



Опублікована стаття буде доступна з 20.06.2025 за посиланням:
<https://archive.liga.science/index.php/universum/issue/view/june2025>

.....

Електронні сертифікати про публікацію та подяки науковим керівникам також будуть доступні з 20 червня. Розсилка замовлених друкованих примірників, сертифікатів та подяк відбудеться з 3 по 10 липня.

З повагою,

Директор Молодіжної наукової ліги
Голова редакційного комітету
ІГОР КОРЕНЮК

ДЯКУЮ ЗА УВАГУ!

Питання?