

Є. М. Крижановський¹В. О. Караваєв¹І. М. Штельмах¹О. О. Войцеховська¹

АВТОМАТИЗАЦІЯ ВИКОРИСТАННЯ ПРИРОДНОМОВНИХ ЗАПИТІВ ДЛЯ КОМПЛЕКСНОГО АНАЛІЗУ СТАНУ ПОВЕРХНЕВИХ ВОД БАСЕЙНУ ПІВДЕННОГО БУГУ

¹Вінницький національний технічний університет

Розглянуто сучасний підхід до комплексного аналізу стану поверхневих вод басейну річки Південний Буг з використанням інтеграції цифрових технологій, необхідних для автоматизації обробки природномовних запитів. Основу запропонованого підходу становить синергія геоінформаційних систем (ГІС), інтелектуальних засобів обробки даних, адаптивного парсингу великих масивів моніторингової інформації та алгоритмів штучного інтелекту — зокрема, великих мовних моделей. Значна увага приділяється автоматизації процесів збору, попередньої обробки, структуризації та візуалізації екологічних даних, що забезпечують якісну підготовку інформаційного середовища для ухвалення управлінських рішень. Розроблений підхід дозволяє здійснювати просторово-часовий аналіз стану водних ресурсів, визначати ключові тенденції забруднення та здійснювати комплексний аналіз стану поверхневих вод басейну Південного Бугу, використовуючи природномовні запити. Апробація підходу здійснена в межах реалізації веб-системи на реальних даних моніторингу стану поверхневих вод басейну Південного Бугу та успішного випробовування цієї системи в напрямку коректного оброблення природномовних запитів.

Використання великих мовних моделей аналізу стану поверхневих вод суттєво спрощує процес формування різноманітних екологічних звітів, класифікацій якості та розв'язання інших прикладних задач аналізу даних моніторингу стану поверхневих вод.

Отримані результати підкреслюють доцільність створення гнучких інформаційних систем для моніторингу стану поверхневих вод, які поєднують можливості просторового аналізу, обробки природної мови та машинного навчання. Це дозволяє забезпечити ухвалення обґрунтованих управлінських рішень та оперативно реагувати на зміни екологічного стану. Запропонований підхід може бути адаптований до інших водних басейнів чи суббасейнів, відкриваючи нові можливості для сталого управління природними ресурсами.

Ключові слова: геоінформаційні системи, моніторинг вод, великі мовні моделі, адаптивний парсинг, аналіз даних.

Вступ

Сучасні методи управління водними ресурсами здебільшого використовують розрізнені дані та не враховують комплексного впливу екологічних факторів. Інтеграція геоінформаційних систем (ГІС), методів системного аналізу та інструментів штучного інтелекту дозволяє об'єднувати, обробляти й аналізувати просторові й моніторингові дані для ефективнішого управління водними ресурсами. *Метою роботи* є розробка сучасного підходу до комплексного аналізу стану поверхневих вод басейну річки Південний Буг, використовуючи інтеграцію цифрових технологій, необхідних для автоматизації природномовних запитів.

Для досягнення поставленої мети необхідно розв'язати такі задачі:

- автоматизувати збір і аналіз даних моніторингу;
- здійснити геопросторову прив'язку даних моніторингу стану вод басейну;
- застосувати сучасні технології штучного інтелекту (ШІ) для обробки природномовних запитів;
- розробити веб-систему на реальних даних моніторингу стану поверхневих вод басейну Південного Бугу та здійснити її випробовування.

Результати цього дослідження сприятимуть підвищенню ефективності моніторингу, поліпшенню управлінських рішень та забезпечать зручні інструменти для роботи з даними.

Аналіз актуальних досліджень

Комплексне дослідження стану поверхневих вод з використанням сучасних цифрових технологій набуває все більшої актуальності в умовах зростаючого антропогенного навантаження та кліматичних змін. Традиційні методи моніторингу, засновані на точкових вимірюваннях і статистичному аналізі, поступово доповнюються новими підходами, які використовують потенціал штучного інтелекту та великих мовних моделей (ВММ).

У роботах [1], [2] підкреслюється важливість інтеграції ГІС з методами машинного навчання для покращення просторової інтерпретації екологічних даних. Особливо перспективним напрямом є використання ВММ, таких як GPT, BERT та їхніх адаптацій, для автоматизованого оброблення неструктурованої інформації — звітів моніторингу, екологічних протоколів, публікацій тощо [3]. Завдяки цьому значно розширюються можливості систематизації знань та виявлення причинно-наслідкових зв'язків у водних екосистемах.

Дослідження [4] демонструють, що поєднання ВММ і нейронних мереж дозволяє автоматизувати аналіз тенденцій забруднення поверхневих вод, зокрема через побудову семантичних моделей, які враховують часову динаміку та контекст. Інтеграція таких підходів у рамках адаптивного парсингу даних спрощує процеси збору й оновлення екологічної інформації, що особливо важливо для великих річкових басейнів, як-от Південний Буг.

До того ж, праці [5], [6] свідчать про ефективність використання ВММ у геоекоекологічному моделюванні для ідентифікації ризикових зон забруднення та підтримки прийняття рішень у сфері водного менеджменту. У цьому контексті великі мовні моделі служать інструментом для інтеграції розрізнених джерел інформації, створення єдиного цифрового середовища знань та підвищення точності прогнозів.

Таким чином, актуальні дослідження підтверджують доцільність застосування ВММ у комплексному аналізі стану поверхневих вод. Їх інтеграція з ГІС, нейронними мережами та алгоритмами аналізу даних формує нову парадигму екологічного моніторингу, що базується на відкритих даних, адаптивності та інтелектуальній автоматизації.

Вибір оптимальних ІТ-інструментів

Для автоматизації обробки природномовних запитів для комплексного аналізу стану поверхневих вод басейну Південного Бугу вибрано низку сучасних інструментів і технологій, які забезпечують ефективну обробку, інтеграцію, аналіз і візуалізацію моніторингових та геопросторових даних (рис. 1).

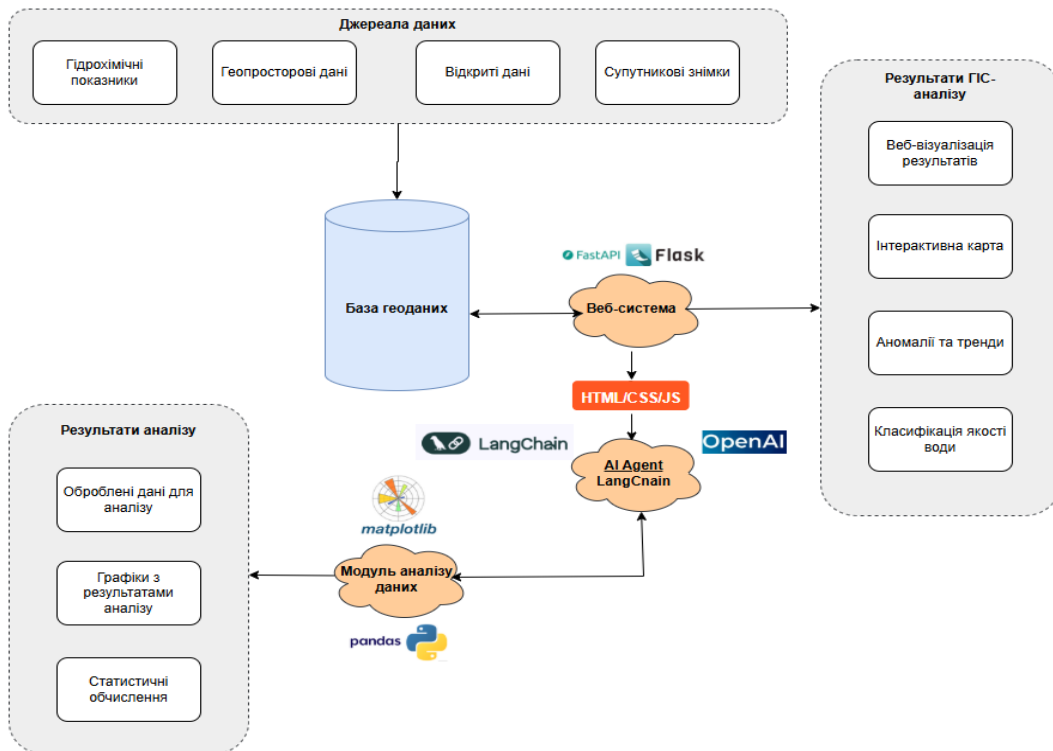


Рис. 1. Вибрані оптимальні інструменти і технології

У сфері веб-розробки для створення зручного інтерфейсу користувача використано HTML5 для побудови структури веб-сторінок, CSS3 для стилізації та реалізації адаптивного дизайну, а також JavaScript (версії ES6+) для реалізації клієнтської логіки. Як серверну платформу вибрано Python, що забезпечує високу швидкодію та простоту інтеграції з бібліотеками для обробки даних. Для розгортання API застосовано FastAPI — сучасний асинхронний фреймворк, який дозволяє швидко створювати веб-сервіси, а також Flask — для реалізації допоміжних веб-компонентів.

Для обробки природномовних запитів користувачів і реалізації елементів штучного інтелекту застосовуються OpenAI API та бібліотека LangChain, що дозволяє інтегрувати мовні моделі у веб-систему та здійснювати логіку природномовної взаємодії.

Обробка табличних і часових даних виконується за допомогою бібліотеки Pandas, а для роботи з геопросторовими об'єктами — GeoPandas, що забезпечує ефективну взаємодію з просторовими форматами та картографічними даними.

Для візуалізації результатів аналізу використано бібліотеку Matplotlib, яка дозволяє будувати графіки, карти й діаграми для відображення динаміки параметрів стану водного середовища.

Автоматизований збір, обробка та структуризація даних моніторингу з просторовою прив'язкою

Джерелом офіційних даних моніторингу стану поверхневих вод є сайт «Моніторинг та екологічна оцінка водних ресурсів України», вибравши необхідні фільтри (період (дату «з» і «до»), басейн Південний Буг та всі доступні показники) можна завантажити дані [7].

Отриманий звіт міститиме детальну інформацію про якість води на основі даних з постів спостереження [8]. Зокрема, будуть отримані концентрації забруднювальних речовин (наприклад, нітратів, фосфатів, важких металів), рівень розчиненого кисню, показники жорсткості води, мінералізації, рН, температура та інші важливі гідрохімічні параметри. Такий звіт дозволить оцінити екологічний стан водойм, порівняти результати з нормативами якості води та виявити потенційно проблемні ділянки для подальшого моніторингу або вжиття коригувальних заходів (рис. 2).

1	Дані моніторингу (за районом річкового басейну або суббасейну)									
2	з 01.03.2000 до 11.03.2025									
3	Район									
4	1. Пост: Балка Ташлик, 6 км, Ташлицьке вдсх., став-охолоджувач ПУ АЕС (міст через Ташлицьке вдсх., південь м. Півден									
5	Значення	Показник								
6		Амоній-іони, мг	Біохімічне спож	Завислі (с	Кисень ро	Нітрат-іон	Нітрит-іон	Сульфат-іс	Фосфат-іс	Хлорид-іон
7	Дата									
8	17.04.2000	0,035	1,3	0	12,98	4,7	0,038	403,2	0,143	146,1
9	04.07.2000	0,031	1,13	10,4	9,89	0,24	0,022	416	0,06	149
10	04.12.2000	0,042	1,2	0	9,63	2,02	0,006	384	0,275	154,5
11	06.02.2001	0,062	2,88	1,6	11,77	4,3	0,004	364,8	0,38	160,1

Рис. 2. Приклад звіту про стан поверхневих вод

Отримані дані містять ключову інформацію про пости спостереження, зокрема назву, географічне розташування, дату відбору проб та значення показників якості води (температура, рН, вміст розчиненого кисню, забруднювальні речовини, тощо). Ці дані є основою для аналізу динаміки зміни екологічного стану водойм, виявлення потенційно небезпечних зон та своєчасного реагування на негативні зміни.

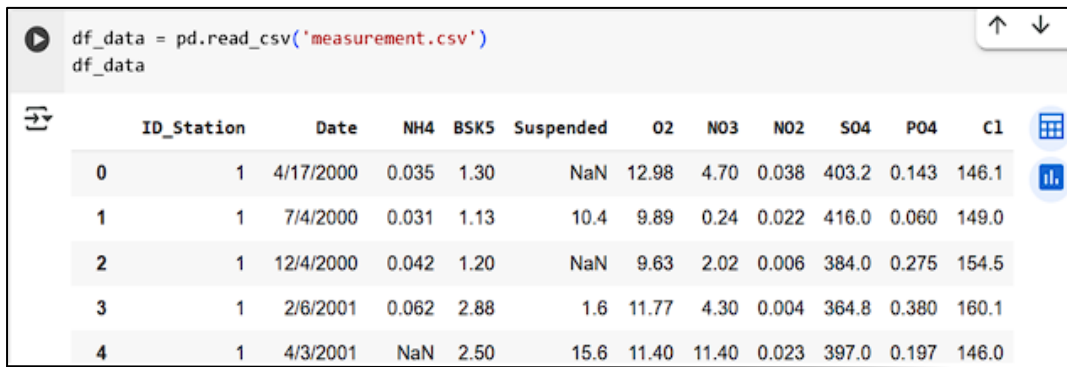
Для формування структурованого набору даних текстовий файл розбивається на рядки, з яких за допомогою регулярних виразів виділяються номери та назви постів. Далі визначаються заголовки показників і обробляються вимірювання: дані розділяються за «;», фіксується дата, значення конвертуються у формат float або позначаються як NaN. У підсумку створюються два DataFrame — для постів спостереження та очищених даних, що використовуються в аналітиці і візуалізації (рис. 3).

```
df_post = pd.read_csv('name_coord_posts.csv')
df_post
```

	post_number	post_name
0	1	Балка Ташлик, 6 км, Ташлицьке вдсх., став-охол...
1	2	Бузький лиман, 0,5 км, скид МКП "Миколаїводока...
2	3	Возсіятське водосховище, 0 км, у межах села В...
3	4	р. Баран, 17 км, м. Жмеринка
4	5	р. Велика Вись, 95 км, с. Лікареве, Новомиргор...
5	6	р. Гнилий Тікич, 41 км, м. Звенигородка, питни...

Рис. 3. Назви постів

Файл measurement містить дані моніторингу якості води — концентрації показників (амоній, кисень, нітрати, фосфати тощо) для кожного поста на відповідну дату (рис. 4).



	ID_Station	Date	NH4	BSK5	Suspended	O2	NO3	NO2	SO4	PO4	Cl
0	1	4/17/2000	0.035	1.30	NaN	12.98	4.70	0.038	403.2	0.143	146.1
1	1	7/4/2000	0.031	1.13	10.4	9.89	0.24	0.022	416.0	0.060	149.0
2	1	12/4/2000	0.042	1.20	NaN	9.63	2.02	0.006	384.0	0.275	154.5
3	1	2/6/2001	0.062	2.88	1.6	11.77	4.30	0.004	364.8	0.380	160.1
4	1	4/3/2001	NaN	2.50	15.6	11.40	11.40	0.023	397.0	0.197	146.0

Рис. 4. Показники якості води

Стовпці post_number і ID_Station використовуються для ідентифікації постів моніторингу: перший — як порядковий номер, другий — як унікальний код станції. Для відображення постів на карті додано координати lat і lon, що дозволяє точно визначити їхнє розташування.

Інтерактивна аналітика даних моніторингу з використанням ШІ

Для геокодування постів використано їхні координати. Розробка інтерактивної карти з використанням GeoPandas та візуалізація даних моніторингу водних ресурсів базується на підходах, описаних у [9]. Пости відображаються на інтерактивній карті у вигляді точкових об'єктів з градацією кольору фону залежно від концентрації забруднювальних речовин.

Дані про якість води доповнюють кожний пост інформацією про рівень забрудненості за наявними показниками якості на основі вимірів шкідливих речовин (рис. 5). Біля кожного маркера на карті відображається анотація з назвою поста та даними про забруднення.

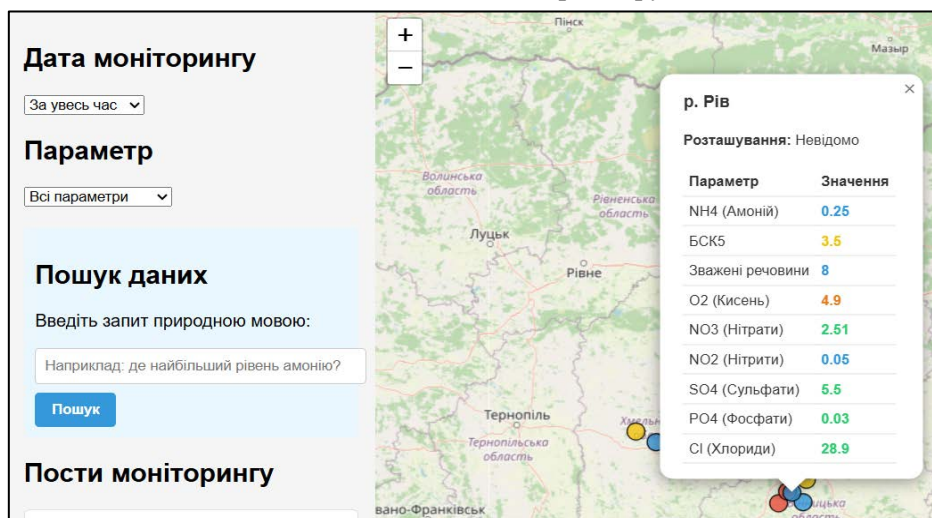


Рис. 5. Інтерфейс веб-системи моніторингу якості води

До карти також додано маркери підприємств — потенційних джерел забруднення, що дозволяє оцінити їхній вплив на якість води. Для кожного посту також передбачена можливість перегляду графіка, який ілюструє динаміку забрудненості за всіма показниками якості води протягом усього періоду спостережень (рис. 6). Графік дозволяє відстежувати динаміку забруднення на постах, виявляючи тенденції поліпшення або погіршення стану води.

Запропоновано підхід до обробки даних спостережень, що використовує можливості великих мовних моделей (LLM) на базі GPT-4, бібліотеки LangChain і pandas. Це дозволяє автоматизувати аналіз великих масивів даних, спрощуючи процес отримання аналітичних висновків з неструктурованих або напівструктурованих джерел.

На рис. 7 показано приклад роботи аналітичної системи, шляхом оброблення запитів природною мовою.

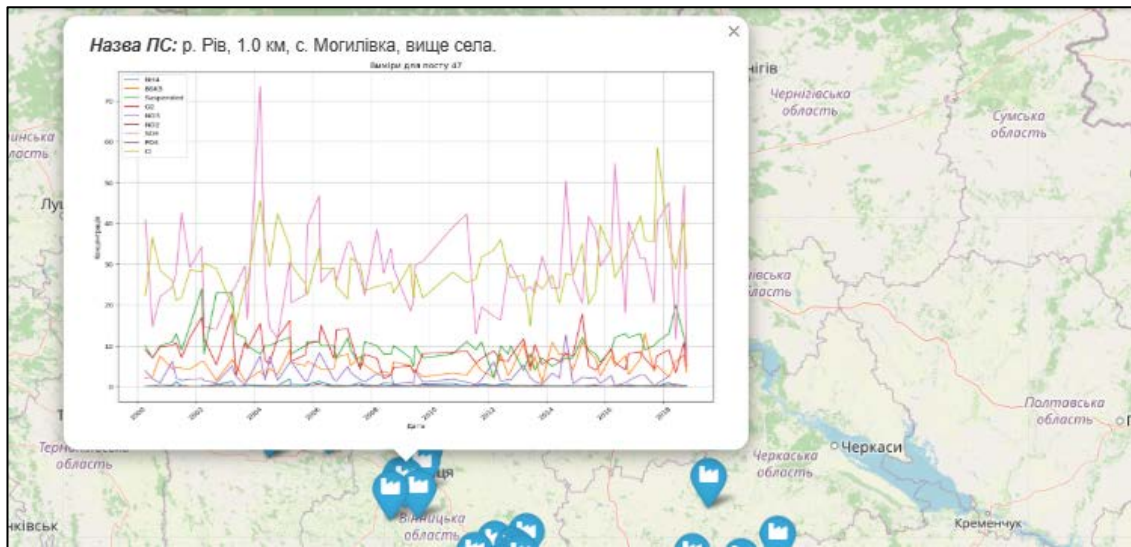


Рис. 6. Графік динаміки зміни показників якості поверхневих вод у вибраному пості спостережень

Запитання: Який самий південний пункт спостереження?
Відповідь: Самий південний пункт спостереження - "Бузький лиман".

Запитання: Яка протяжність ділянки спостереження?
Відповідь: Максимальна протяжність ділянки спостереження становить 773 кілометри.

Запитання: Який найбільший рівень BSK5 спостерігався і коли, на якому посту?
Відповідь: Найвищий рівень BSK5, який спостерігався, становив 50.9. Це відбулося на посту з ID 41 24 листопада 2003 року.

Запитання: В якому місяці спостерігається мінімальний і максимальний рівень BSK5?
Відповідь: Мінімальний рівень BSK5 спостерігається у січні, а максимальний - у березні.

Рис. 7. Приклади роботи аналітичної системи, яка обробляє запити природною мовою

Програма працює з двома джерелами даних: перше містить координати та назви пунктів спостереження, друге — результати екологічних вимірювань. На основі цих даних створюється агент, що забезпечує інтерактивну взаємодію з користувачем через природномовні запити. Функціональну схему роботи такого ШІ-агента показано на рис. 8.

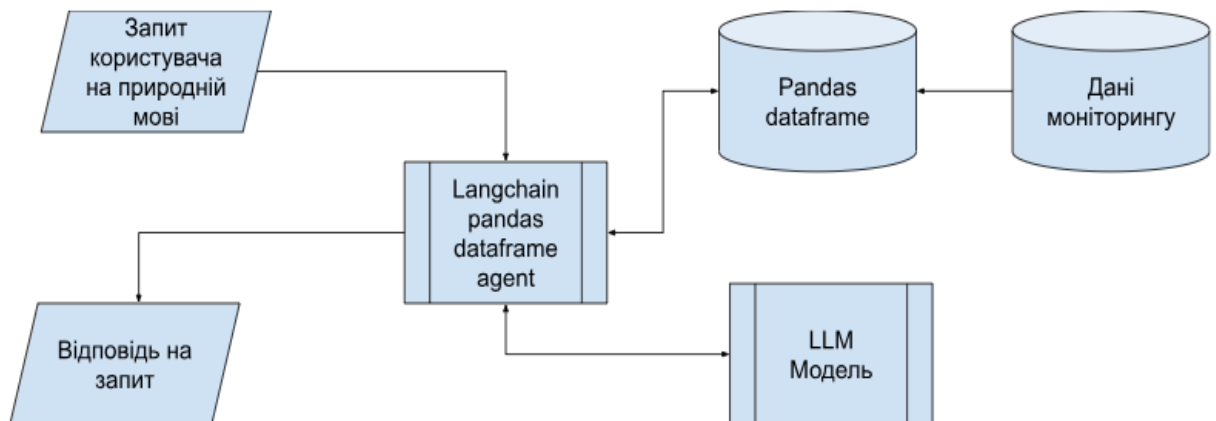


Рис. 8. Функціональна схема роботи ШІ-агента

Моніторингові дані зчитуються з бази даних у датафрейми за допомогою Pandas. LangChain створює ШІ-агента на базі цих датафреймів і GPT-4 [10]. Користувач формулює запит природною мовою, LLM аналізує його, визначає потрібні обчислення та виконує їх локально, не передаючи великі обсяги даних. Результат повертається у вигляді зрозумілої відповіді. Такий підхід усуває потребу у знаннях програмування чи статистики — достатньо сформулювати запит звичайною мовою (рис. 9).

Шаблони пошуку — система розпізнає такі типи запитів:

- максимальні значення: ['максимум', 'найбільш', 'найвищ', 'найгірш'],
- мінімальні значення: ['мінімум', 'найменш', 'найнижч'],
- порівняння "більше": ['більше', 'вище', 'понад', 'перевищує'],
- порівняння "менше": ['менше', 'нижче', 'до', 'не перевищує'],
- якість води: ['найкраща', 'найгірша'],
- забруднення: ['найгірш', 'найбрудніш', 'забруднен'].

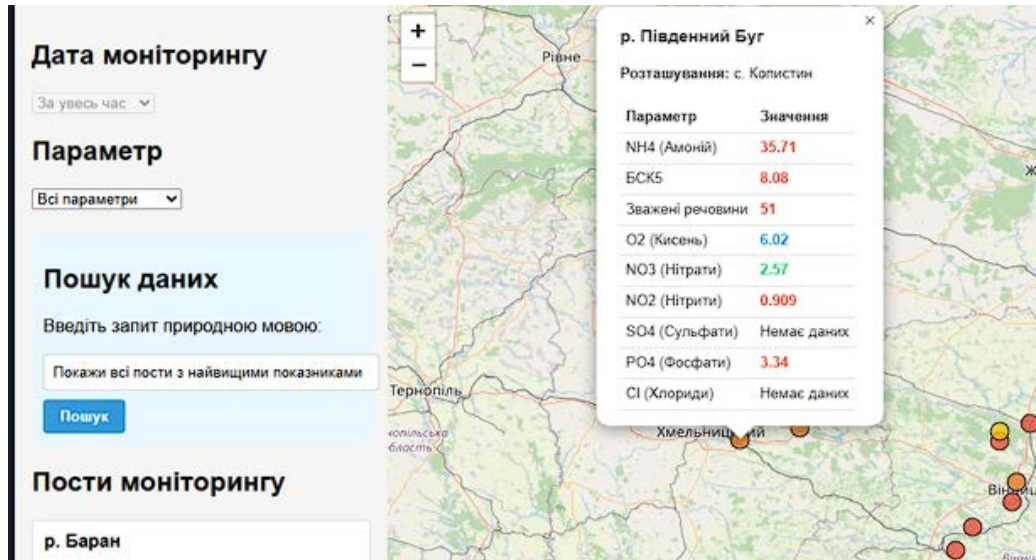


Рис. 9. Приклад запити ШІ-агента та результат його виконання

Таким чином, розроблена веб-система демонструє ефективну інтеграцію LLM-моделі з Pandas на основі бібліотеки LangChain, як середовища для високопродуктивної обробки даних. Вона ілюструє перспективний напрям розвитку аналітичних систем, де ключову роль відіграє не написання коду або здійснення розрахунків, а лише формування опису даних або розрахунків, які потрібні користувачу.

Запропонований підхід має низку обмежень, за яких можлива його коректна практична реалізація та надійне функціонування системи, створеної на базі цього підходу:

- якість вхідних даних моніторингу стану поверхневих вод — наявність даних регулярних спостережень стану поверхневих вод та коректних географічних координат постів спостережень;
- коректність (однозначність) природомовних запитів, які задає користувач системи.

Висновки

Проведене дослідження свідчить про перспективність поєднання сучасних цифрових технологій, таких як геоінформаційні системи, LLM, адаптивний парсинг для розв'язання задач моніторингу та аналізу стану поверхневих вод басейну річки Південний Буг.

Запропоновано підхід до комплексного аналізу стану поверхневих вод басейну річки Південний Буг за використання інтеграції цифрових технологій, необхідних для автоматизації природномовних запитів. Основу запропонованого підходу становить синергія ГІС, інтелектуальних засобів обробки даних, адаптивного парсингу великих масивів моніторингової інформації та алгоритмів штучного інтелекту — зокрема, великих мовних моделей. Значну увагу приділено автоматизації процесів збору, попередньої обробки, структуризації та візуалізації екологічних даних, що забезпечить якісну підготовку інформаційного середовища для ухвалення управлінських рішень.

До того ж, ГІС-аналіз у поєднанні з методами системного підходу дозволив візуалізувати геопросторові зв'язки між джерелами забруднення, рівнями забрудненості, сезонними змінами та іншими факторами, які впливають на якість поверхневих вод.

Також, розроблений підхід дозволяє здійснювати просторово-часовий аналіз стану водних ресурсів, визначати ключові тенденції забруднення та здійснювати комплексний аналіз стану поверхневих вод басейну Південного Бугу шляхом використання природномовних запитів. Апробація

підходу здійснена в межах реалізації веб-системи на реальних даних моніторингу стану поверхневих вод басейну Південного Бугу та успішного випробовування цієї системи в для коректного оброблення природномовних запитів.

Отримані результати підкреслюють доцільність створення гнучких інформаційних систем для моніторингу стану поверхневих вод, які поєднують можливості просторового аналізу, обробки природної мови та машинного навчання. Це дозволяє забезпечити ухвалення обґрунтованих управлінських рішень та оперативно реагувати на зміни екологічного стану. Запропонований підхід може бути адаптований до інших водних басейнів чи суббасейнів, відкриваючи нові можливості для сталого управління природними ресурсами.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

- [1] H. Zhang, "AI and Big Data in Water Environments," *ACS ES&T Water*, vol. 2, issue 6, pp.904-906, 2022. <https://doi.org/10.1021/acsestwater.2c00203>.
- [2] D. B. Olawade, et al., "Artificial intelligence in environmental monitoring: Advancements, challenges, and future directions," *Hygiene and Environmental Health Advances*, vol. 12, 100114, 2024. <https://doi.org/10.1016/j.heha.2024.100114>.
- [3] В. Б. Мокін, М. А. Гораш, Є. М. Крижановський, і Т. Є. Вуж, «Інформаційна інтелектуальна технологія автоматизованої геоприв'язки екологічної текстової природно-мовної інформації», *Наукові праці ВНТУ*, № 4, 2022. [Електронний ресурс]. Режим доступу: <https://praci.vntu.edu.ua/index.php/praci/article/view/624>.
- [4] European Environment Agency. Europe's State of Water 2024: The need for improved water resilience. EEA Report no. 07/2024. [Online]. Available: <https://www.eea.europa.eu/en/analysis/publications/publications/europes-state-of-water-2024>.
- [5] M. Landt-Hayen, et al., "A climate index collection based on model data," *Environmental Data Science*, vol. 2, pp. 1-13, 2023. <https://doi.org/10.1017/eds.2023.5>.
- [6] Кабінет Міністрів України. Постанова від 19 вересня 2018 р. № 758 (зі змінами № 1071 від 06.09.2024) «Про затвердження порядку здійснення державного моніторингу вод». [Електронний ресурс]. Режим доступу: <https://zakon.rada.gov.ua/laws/show/1071-2024-%D0%BF#n25>.
- [7] Веб-система «Моніторинг та екологічна оцінка водних ресурсів України». 2024. [Електронний ресурс]. Режим доступу: <http://monitoring.davr.gov.ua/EcoWaterMon/GDKMap/Index>.
- [8] C. Kermorvant, et al., "Understanding links between water-quality variables and nitrate concentration in freshwater streams using high-frequency sensor data," *Plos one*, vol. 18, issue 6, e0287640, 2023. <https://doi.org/10.1371/journal.pone.0287640>.
- [9] В. Б. Мокін, Є. М. Крижановський, і М. П. Боцула, *Інформаційна технологія інтегрування математичних моделей у геоінформаційні системи моніторингу поверхневих вод*, моногр. Вінниця, Україна: ВНТУ, 2011. [Електронний ресурс]. Режим доступу: https://ir.lib.vntu.edu.ua/bitstream/handle/123456789/1943/Mokin_ITIntegrMatModelUGeoInformSistMonitPoverhVod394.pdf?sequence=1&isAllowed=y.
- [10] Y. Tian, et al., "Advancing Large Language Models for Spatiotemporal and Semantic Association Mining of Similar Environmental Events," *Transactions in GIS*, vol. 29, e13282, 2025. <https://doi.org/10.1111/tgis.13282>.

Рекомендовано кафедрою системного аналізу та інформаційних технологій ВНТУ

Стаття надійшла до редакції 27.06.2025

Крижановський Євгеній Миколайович — канд. техн. наук, доцент, доцент кафедри системного аналізу та інформаційних технологій, e-mail: kruzhan@gmail.com ;

Каравасєв Вадим Олександрович — студент факультету інтелектуальних інформаційних технологій та автоматизації, e-mail: karavaevvadim1999@gmail.com ;

Штельмах Ігор Миколайович — канд. техн. наук., асистент кафедри системного аналізу та інформаційних технологій, e-mail: igor.shtelmakh@vntu.edu.ua ;

Войцеховська Ольга Олександрівна — д-р філософії, старший викладач кафедри системного аналізу та інформаційних технологій, e-mail: olgav1085@gmail.com .

Вінницький національний технічний університет, Вінниця

Ye. M. Kryzhanovskiy¹
V. O. Karavaev¹
I. M. Shtelmakh¹
O. O. Voitsekhovska¹

Automation of the Use of Natural Language Queries for Comprehensive Analysis of the State of Surface Waters in the Southern Bug River Basin

¹Vinnitsia National Technical University

The article considers a modern approach to a comprehensive analysis of the state of surface waters in the Southern Bug River basin by integrating digital technologies necessary for the automation of natural language queries. The basis of the proposed approach is the synergy of geographic information systems (GIS), intelligent data processing tools, adaptive parsing of large arrays of monitoring information and artificial intelligence algorithms - in particular, large language models. Considerable attention is paid to the automation of the processes of collection, pre-processing, structuring and visualization of environmental data, which ensures high-quality preparation of the information environment for making management decisions. The developed approach allows for a spatio-temporal analysis of the state of water resources, identification of key pollution trends and a comprehensive analysis of the state of surface waters in the Southern Bug basin by using natural language queries. The approach was tested within the framework of the implementation of a web system on real data for monitoring the state of surface waters of the Southern Bug basin and the successful testing of this system in the direction of correct processing of natural language queries.

The use of large language models for the analysis of the state of surface waters significantly simplifies the process of forming various environmental reports, quality classifications and solving other applied problems of analyzing data on the state of surface waters monitoring.

The results obtained emphasize the feasibility of creating flexible information systems for monitoring the state of surface waters, which combine the capabilities of spatial analysis, natural language processing and machine learning. This allows for making informed management decisions and promptly responding to changes in the environmental state. The proposed approach can be adapted to other water basins or sub-basins, opening up new opportunities for sustainable management of natural resources.

Keywords: geographic information systems, water monitoring, large language models, adaptive parsing, data analysis.

Kryzhanovskiy Yevhenii M. — Cand. Sc. (Eng.), Associate Professor, Associate Professor of the Chair of System Analysis and Information Technologies, e-mail: kruzhan@gmail.com ;

Karavaev Vadim O. — Student of the Department of Intelligent Information Technologies and Automation, e-mail: karavaevvadim1999@gmail.com ;

Shtelmakh Ihor M. — Cand. Sc. (Eng.), Assistant of the Chair of System Analysis and Information Technologies, e-mail: igor.shtelmakh@vntu.edu.ua ;

Voitsekhovska Olha O. — PhD, Senior Lecturer of the Chair of System Analysis and Information Technologies, e-mail: olgav1085@gmail.com