Технічні деталі аналізу

1. Відбір релевантних повідомлень

На першому етапі модель GPT аналізувала повний текст кожного повідомлення, щоб визначити, чи згадуються в ньому результати соціологічного опитування або дослідження. Використовувався промпт із чіткими правилами та двома можливими відповідями (“Так” або “Ні”). Повідомлення, у яких згадувалося лише абстрактне, майбутнє або неконкретне опитування, відфільтровувалися.

Модель GPT: gpt-4o mini.

Промпт:

> *Persona: Ти — уважний аналітик, що спеціалізується на перевірці джерел.*
>
> *Завдання: Проаналізуй наступний текст і визнач, чи згадуються в ньому результати *конкретного*, *вже проведеного* соціологічного опитування або дослідження (хоча б одного).*
>
> *Контекст (Правила): 1. Якщо згадується абстрактне опитування ('соціологи кажуть', 'опитування показують'), відповідай 'Ні'. 2. Якщо згадується майбутнє опитування ('ми плануємо опитати'), відповідай 'Ні'. 3. Якщо згадується конкретне опитування (є назва, організатор, дата або чіткі результати), відповідай 'Так'.*
>
> *Формат: Відповідай лише одним словом: 'Так' або 'Ні'.*
>
> *Текст повідомлення:*
>
> *[txt]*

2. Виділення організації чи організацій, що проводили або замовляли опитування, та короткого опису теми опитування

Для повідомлень, що пройшли попередню фільтрацію, застосовувався один комбінований промпт для одночасного вилучення назв виконавця та замовника опитування, а також короткого опису теми опитування.

Модель GPT: gpt-4o mini.

Промпт:

> *Ти — аналітичний асистент. Твоє завдання — вилучити з тексту дані про походження соціологічного дослідження.*
>
> *КРОК 1. Знайди організації, які відіграють одну з двох ролей:*
>
> *1. ВИКОНАВЕЦЬ (хто провів опитування, збирав дані, соціологічна служба/центр).*
>
> *2. ЗАМОВНИК (хто ініціював, фінансував або на чиє замовлення робили опитування).*
>
> *Критерії пошуку:*
>
> *- Шукай фрази: 'на замовлення...', 'проведено компанією..., 'дослідження центру...', 'спільно з...'.*
>
> *- Якщо організація лише ОПРИЛЮДНИЛА новину (наприклад, ЗМІ, телеканал, новинний сайт), але не є замовником чи виконавцем дослідження — НЕ включай її.*
>
> *- Якщо знайдено більше двох організацій (наприклад, альянс замовників і виконавців) — обери дві найважливіші (першоджерела).*
>
> *КРОК 2. Сформулюй тему опитування одним реченням.*

ФОРМАТ ВИВОДУ (суворо два рядки):

Хто проводив: <Назва Виконавця>; <Назва Замовника>

Тема: <Тема опитування>

Якщо в тексті не вказано ні виконавця, ні замовника, поверни: 'Хто проводив:'

Більше ніяких пояснень."

[text]

3. Нормалізація та уніфікація назв організацій за словником

Застосовувалися текстові правила для видалення зайвих слів, символів, регістрів тощо. Для уніфікації назв організацій використовувався попередньо сформований словник із найпопулярнішими організаціями, їх найпоширенішими варіантами написання та стандартизованими відповідниками.

4. Групування та уніфікація назв організацій поза словником

Цей етап передбачав кілька послідовних кроків для усіх неуніфікованих за словником назв організацій:

4.1. Початкове групування на основі входження рядків

Унікальні назви організацій (окрім стоп-слів типу “центр”, “інститут”) групувалися, якщо одна назва містилася в іншій. Групи сортувалися за довжиною назв всередині групи та за першим елементом між групами.

4.2: Ітеративне попарне порівняння та уніфікація за допомогою GPT

Відсортований список згрупованих назв проходив ітеративну обробку. Пари сусідніх назв подавалися GPT для визначення, чи позначають вони ту саму організацію, і якщо так, то яка з назв є коротшою/коректнішою.

Модель GPT: gpt-4o.

Промпт:

> *Порівняйте наступні дві назви організацій:*
>
> *1) '[prev_name]'*
>
> *2) '[next_name]'*
>
> *Якщо ці назви позначають ту ж саму організацію, і одна з них є коротшою, поверніть коротшу назву.*
>
> *Якщо це різні організації, поверніть першу назву.*

4.3: Фінальне групування топ-150 назв за допомогою GPT

Було сформовано список 150 найчастіше згадуваних назв організацій поза словником. Список подавався GPT з проханням упорядкувати його так, щоб дублікати йшли один за одним.

Модель GPT: gpt-4o.

Промпт:

> *Ось список 150 назв організацій, впорядкованих за частотою:*
>
> *[top_150_names]*
>
> *Будь ласка, проаналізуйте цей список і, якщо знайдете дублікати (назви, що позначають ту ж організацію), упорядкуйте їх так, щоб повторювані назви йшли один за одним.*
>
> *Вихідний формат: список назв (без нумерації), по одному рядку.*

4.4. Кластеризація та вибір представницької назви GPT

Очищений та впорядкований список назв подавався GPT для кластеризації та вибору найкоротшої представницької назви для кожного кластера.

Модель GPT: gpt-4o.

Промпт:

> *Ось список назв організацій:*
>
> *[good_lines]*
>
> *Групуйте всі варіанти назв однієї організації в кластери. Для кожного кластера оберіть найкоротшу назву як представницьку і виведіть рядок у форматі:*
>
> *представницька (варіант1, варіант2, ...)*
>
> *Якщо кластер містить лише одну назву, виведіть її без дужок.*

Відповідь GPT парсилася для створення таблиці відповідності “варіант назви -> канонічна назва”. Ця таблиця використовувалася для оновлення назв організацій, що не були уніфіковані за допомогою словника.

4.5. Перетворення

Уніфіковані назви організацій приводились до коректного вигляду. Зокрема, видалялася друга назва, якщо обидві були ідентичними.

5. Аналіз якості подання результатів опитувань

Для кожного повідомлення за допомогою GPT ми визначали наявність 10 типів методологічної інформації (замовник, виконавець, дати опитування, генеральна сукупність, розмір вибірки, метод вибірки, похибка, зважування, метод опитування, текст запитання).

Модель GPT: gpt-4o mini.

Промпт:

Ти — експерт з методології соціологічних досліджень. Твоє завдання — проаналізувати текст прес-релізу або звіту та визначити, чи містить він інформацію, що відповідає стандартам AAPOR (American Association for Public Opinion Research).

Для кожного з 10 пунктів нижче напиши 'так', якщо інформація наявна в тексті, або 'ні', якщо вона відсутня. Використовуй наведені визначення AAPOR для прийняття рішення:

1. Замовник дослідження (Sponsor / Funding Source):

- Шукай: чітку вказівку на те, хто фінансував або замовив опитування (наприклад, 'на замовлення...', 'за підтримки фонду...', 'фінансується урядом...').

- AAPOR Definition: Name the sponsor of the research. If the original source of funding is different than the sponsor, this source will also be disclosed.

2. Виконавець дослідження (Who Conducted the Research):

- Шукай: назву організації, яка безпосередньо проводила опитування, збирала та обробляла дані (соціологічний центр, дослідницька агенція).

- AAPOR Definition: Name the party(ies) who conducted the research.

3. Дати проведення опитування (Dates of Data Collection):

- Шукай: часові межі збору даних.

- Критерій 'так': вказано конкретні дні (наприклад, '10-15 березня') АБО вказано місяць і рік (наприклад, 'у березні 2024 року').

- Критерій 'ні': вказано ЛИШЕ рік (наприклад, 'опитування 2023 року') або дати відсутні.

- AAPOR Definition: Disclose the dates of data collection.

4. Генеральна сукупність (Population Under Study):

- Шукай: кого саме репрезентує вибірка.

- Критерій 'так':

a) Для загальних опитувань: достатньо загальної вказівки на громадянство чи країну (наприклад, 'опитано українців', 'населення України', 'більшість громадян вважає').

b) Для специфічних груп: вказано конкретні характеристики (наприклад, 'ВПО', 'лікарі', 'молодь 18-25 років').

- AAPOR Definition: Researchers will be specific about the decision rules used to define the population (location, age, other social or demographic characteristics).

5. Розмір вибірки (Sample Sizes):

- Шукай: кількість опитаних респондентів (загальна або по групах).

- AAPOR Definition: Provide sample sizes for each mode of data collection (for surveys include sample sizes for each frame, list, or panel used).

6. Метод вибірки (Method Used to Generate and Recruit the Sample):

- Шукай: як відбирали респондентів (випадкова вибірка, квотна, 'снігова куля', панель тощо).

- AAPOR Definition: Explicitly state whether the sample comes from a frame selected using a probability-based methodology or if the sample was selected using non-probability methods (opt-in, volunteer). Describe any use of quotas.

7. Похибка вибірки (Precision of the Results):

- Шукай: згадки про статистичну похибку, маржу помилки (margin of error) або точність.

- AAPOR Definition: For probability sample surveys, report estimates of sampling error. Reports of non-probability sample surveys will only provide measures of precision if they are defined and accompanied by a detailed description.

8. Застосування вагових коефіцієнтів (How the Data Were Weighted):

- Шукай: інформацію про зважування даних (weighting) для корекції вибірки.

- AAPOR Definition: Describe how the weights were calculated, including the variables used and the sources of the weighting parameters.

9. Метод проведення опитування (Method(s) and Mode(s) of Data Collection):

- Шукай: метод контакту (телефон, CATI, онлайн-панель, особисте інтерв'ю, CAPI, пошта).

- AAPOR Definition: Include a description of all mode(s) used to contact participants or collect data or information (e.g., CATI, CAPI, ACASI, IVR, mail, Web).

10. Текст запитання (Measurement Tools/Instruments):

- Шукай: точне формулювання запитання, яке ставили респондентам.

- AAPOR Definition: The exact wording and presentation of any measurement tool from which results are reported.

ФОРМАТ ВІДПОВІДІ (суворо дотримуйся порядку):

Замовник дослідження: [так/ні]

Виконавець дослідження: [так/ні]

Дати проведення опитування: [так/ні]

Генеральна сукупність: [так/ні]

Розмір вибірки: [так/ні]

Метод вибірки: [так/ні]

Похибка вибірки: [так/ні]

Застосування вагових коефіцієнтів: [так/ні]

Метод проведення опитування: [так/ні]

Текст запитання: [так/ні]

Текст повідомлення:

[text]

6. Визначення рівня дотримання стандартів розкриття інформації про опитування

На основі 10 елементів методологічної інформації, зібраних на попередньому кроці, було сформовано Індекс дотримання стандартів розкриття інформації про опитування. Концепт містить три виміри: “джерельна доброчесність”, “статистична прозорість” та “інтерпретаційна точність”. Кожен вимір, у свою чергу, має індикатори з числа визначених на попередньому етапі.

“Джерельна доброчесність” — наявність згадок про замовника та виконавця.

“Статистична прозорість” — наявність згадок про генеральну сукупність, розмір та метод вибірки, похибку та зважування.

“Інтерпретаційна точність” — наявність згадок про дати та метод опитування, а також текст запитань.

Кожен індикатор було перекодовано з “так/ні” у “1/0” відповідно. Для кожного виміру обчислюється частка заповнених індикаторів: сума значень індикаторів ділиться на їхню кількість у цьому вимірі. У результаті формуються три окремі показники — по одному для кожного виміру. Загальний індекс визначається як середнє арифметичне цих трьох показників, оскільки усі виміри мають однакову вагу.

Рівні дотримання були категоризовані: низький [0-0.3); середній [0.3-0.6); високий [0.6-1].

7. Класифікація теми дослідження (за закритим переліком)

Тематика кожного опитування була класифікована за одним із 17 попередньо визначених напрямів з дефініціями (див. *Таблиця 1. Тематика опитувань та їх визначення*). На вхід моделі для класифікації подавався короткий опис теми, вилучений на кроці 2, а не повний текст повідомлення.

Модель GPT: gpt-4o mini.

Промпт:

> *Classify the survey's primary topic from these categories:*
>
> *[список_тем_через_крапку_з_комою] .*
>
> *Definitions:*
>
> *[список_визначень_тем_через_крапку_з_комою] .*
>
> *Survey description:*
>
> *[text_from_survey_topic_column]*
>
> *Return only the best-matching topic or 'None'.*

8. Ідентифікація висвітлення однакових опитувань

Опитування ідентифікувалися за трьома критеріями: однакова тема; публікація новин про опитування в межах 5 днів та наявність щонайменше однієї спільної організації. Об’єднання у групу здійснювалося за умови наявності принаймні двох елементів, що відповідали критеріям.

Таблиця 1. Тематика опитувань та їх визначення

Категорія	Що охоплює
Довіра до політиків та інститутів, і врядування	Довіра до органів влади, політичних лідерів, державних інституцій, політичні настрої, врядування.
Зовнішні відносини / Міжнародні справи	Міжнародна дипломатія, зовнішня політика, відносини між країнами, геополітика, міжнародні організації.
Військо та оборона	Збройні сили, національна безпека, бойові дії, військова політика, ставлення до армії.
Економічний та бізнес-клімат	Економічні тенденції, бізнес-середовище, інвестиції, підприємництво, ринок.
Зайнятість і ринок праці	Рівень зайнятості, ринок праці, безробіття, умови праці, трудові права.
Соціальні питання та добробут	Бідність, нерівність, соціальний захист, демографія, вразливі групи, якість життя.
Охорона здоров’я	Медичні послуги, здоров’я населення, інфраструктура, страхування, захворювання.
Освіта	Школи, університети, якість освіти, грамотність, політика в освіті, інфраструктура.
Інфраструктура та міське планування	Транспорт, дороги, житло, розвиток міст, комунальні послуги, просторове планування.
Громадянська участь	Участь у виборах, громадська активність, волонтерство, взаємодія з владою.
Культура та ідентичність	Національна ідентичність, традиції, мова, історична пам’ять, культурні практики.

Правопорядок і судочинство	Злочинність, правоохоронна система, суди, безпека громадян, поліція, тюрми.
Довкілля та клімат	Екологія, зміна клімату, забруднення, охорона природи, стале використання ресурсів.
Медіа та інформація	Медіа, довіра до інформації, споживання новин, дезінформація, медіаграмотність.
Технології	Технології, цифровізація, інтернет, кібербезпека, інновації, діджитал-інфраструктура.
Енергетика	Політика в енергетиці, джерела енергії, ринки палива, відновлювана енергетика.
Міграція	Міграція, біженці, інтеграція, політика щодо іммігрантів, громадська думка про мігрантів.