DOI: 10.34185/1991-7848.2026.01.21
УДК 004.8

М.О. Березюк, А.І. Гуда

МЕТОД ПОБУДОВИ КРИЗОВО-КОНТЕКСТНОГО ДАТАСЕТУ ДЛЯ ВЕРИФІКАЦІЇ ADAPTIVE IRM

Анотація. *Ця робота присвячена не експериментальному підтвердженню ефективності Adaptive IRM, а побудові спеціалізованого кризово-контекстного датасету, який робить таку перевірку можливою в коректній постановці. У статті запропоновано метод перетворення кризових повідомлень із HumAID у пари виду «абстрактний запит – кризово-залежна відповідь», де питання навмисно очищується від прямих маркерів лиха, а правильна інтерпретація потребує відновлення прихованого контексту події. Такий дизайн відрізняється від переважних у crisis informatics задач tweet-level classification, informativeness detection, humanitarian categorization і multimodal crisis annotation, для яких призначені HumAID, CrisisBench, AIDR, TREC-IS і CrisisMMD [1, 2, 3, 4, 5, 6]. У результаті роботи сформовано датасет обсягом 41 152 записи за п'ятьма категоріями кризових подій; під час генерації питань використовувалася схема primary generation -> retry generation -> fallback, причому fallback було задіяно у 1 432 випадках, що становить 3.48% корпусу. Як наступний етап пропонуються формалізована ручна валідація, автоматична retrieval-style перевірка семантичної узгодженості, event-disjoint split на рівні подій HumAID, реалізація Adaptive IRM і порівняння LLM-baseline, LLM+Adaptive IRM, RAG і PEFT-baselines із розширеним набором автоматичних і ручних метрик [7, 8, 9, 10, 11, 12, 13, 14, 15].*

Ключові слова: *кризово-контекстний датасет, великі мовні моделі, прихована контекстна адаптація, Adaptive IRM, question generation, crisis informatics, HumAID, генерація запитань, валідація датасетів.*

Постановка проблеми

Соціальні медіа давно стали одним із ключових джерел даних для crisis informatics, оскільки під час надзвичайних ситуацій користувачі публікують оперативні відомості про руйнування, потреби, переміщення, інфраструктурні збої, пошук допомоги та локальну обстановку. Водночас саме це середовище

створює високе навантаження на автоматичні методи аналізу: повідомлення короткі, шумні, контекстно насичені й не завжди придатні для прямого вилучення actionable information. Тому задача не зводиться до простого розпізнавання тексту: необхідні способи враховувати прихований контекст події, який визначає, як саме має інтерпретуватися один і той самий загальний запит [1].

Актуальність такої постановки посилюється з появою великих мовних моделей. Сучасні огляди з LLM у disaster management показують, що ці моделі вже застосовуються для фільтрації за релевантністю, класифікації, вилучення інформації, summarization і підтримки рішень на різних фазах disaster cycle. Водночас той самий огляд підкреслює відкриті проблеми надійності, доменної адаптації, відтворюваності та якісного оцінювання в кризовому середовищі, де контекст не завжди заданий явно, а помилка відповіді може бути змістовно значущою [7].

На цьому тлі запропонована стаття розв'язує підготовчу, але методологічно важливу задачу: формує корпус для перевірки того, чи може модель коректно відновлювати кризовий контекст не з прямої згадки лиха в питанні, а з окремого контекстного сигналу та семантичного зв'язку з відповіддю. Саме тому центральним науковим результатом тут є датасет і його валідаційна рамка, а не вже реалізована модифікація базової LLM.

Аналіз останніх досліджень і публікацій

Наявні кризові корпуси та постановки

Одним із найбільш релевантних джерел для цієї роботи є HumAID — корпус, що містить близько 77 тисяч вручну розмічених твітів, відібраних із пулу приблизно 24 мільйонів повідомлень, пов'язаних із 19 надзвичайними подіями 2016-2019 років. Автори HumAID розглядають цей ресурс як основу для навчання та порівняння моделей, що працюють із кризовими повідомленнями на рівні окремих твітів, насамперед у задачах класифікації та аналізу гуманітарно значущої інформації [2].

Подібну дослідницьку лінію розвиває CrisisBench. У цій роботі об'єднано вісім розмічених кризових джерел, на основі яких сформовано вибірки обсягом 166.1 тис. твітів для задачі informativeness classification і 141.5 тис. твітів для humanitarian classification. Таким чином, CrisisBench орієнтований на зіставне

ISSN-print 1991-7848
ISSN-online 2707-9457

навчання й оцінювання класифікаторів кризових повідомлень, а не на перевірку здатності моделей адаптувати відповідь до прихованого кризового контексту [3].

AIDR також належить до класу систем, орієнтованих на автоматичну обробку й класифікацію crisis-related microblog communications. Платформу було розроблено для швидкого налаштування користувацьких категорій та автоматичного розподілу повідомлень за ними в умовах надзвичайних ситуацій. TREC-IS, своєю чергою, формує постановку навколо пошуку actionable information у потоках соціальних медіа, використовуючи інформаційні типи та пріоритизацію повідомлень для підтримки response scenarios. CrisisMMD розширює кризовий аналіз завдяки мультимодальним даним, однак його основні анотаційні задачі також залишаються пов'язаними з класифікацією: informative vs. not informative, humanitarian categories і damage severity assessment [4, 5, 6].

Отже, наявні кризові корпусні ресурси послідовно розв'язують задачі класифікації повідомлень, фільтрації, категоризації, пріоритизації та мультимодальної анотації. Ці напрями є фундаментальними для crisis informatics, однак вони не покривають окрему постановку, у якій користувацький запит залишається абстрактним, а релевантна відповідь визначається прихованим кризовим контекстом.

LLM, контекстна адаптація та PEFT

Зв'язок роботи із сучасною літературою щодо LLM та адаптації формулюється через два взаємодоповнювальні напрями. Перший — це disaster-oriented LLM research, де огляд 2025 року щодо LLM у disaster management фіксує зростання прикладних робіт, але також повідомляє про брак стандартизованих оцінювальних рамок, придатних датасетів і акуратної доменної адаптації [7]. Другий — це література з parameter-efficient adaptation, яка показує, що поведінку великих моделей можна предметно перебудовувати без повного fine-tuning. Сюди належать adapters, prefix tuning, LoRA, IA3 і споріднені методи [9, 16, 17, 18, 19].

Саме в цьому контексті Adaptive IRM слід позиціонувати як спеціалізований механізм прихованої контекстної realignment/adaptation, а не як довільну надбудову. Концептуально такий модуль перебуває поруч із PEFT-

підходами за метою — адаптувати поведінку моделі за обмежених ресурсів, — але відрізняється тим, що в розглянутій постановці важливим є не лише перенесення в домен, а відновлення релевантного кризового контексту за абстрактним користувацьким питанням. Для експериментального фону також доречно згадати RAG як сильний baseline зовнішнього контекстного підсилення й AdaptEval як свідчення того, що здатність LLM адаптуватися до домену вже оцінюється як окрема дослідницька проблема [8, 10].

Генерація питань та оцінювання

Для генерації питань стаття спирається не лише на практичну евристику, а й на корпусний напрям question generation. Сучасний огляд із Neural Question Generation систематизує типи входів, методи, benchmark-дані та evaluation metrics і тим самим дає коректну рамку для пояснення, чому автоматична генерація питань може використовуватися як інструмент перетворення вихідного тексту на перевірювану пару «питання — відповідь» [20].

Водночас література з QG-оцінювання прямо попереджає, що reference-based metrics є недостатніми, особливо коли на один контекст припадає лише одне референсне питання. QGEval вводить семивимірну рамку оцінювання — fluency, clarity, conciseness, relevance, consistency, answerability і answer consistency — та показує, що наявні automatic metrics часто погано узгоджуються з human judgments. Робота про те, що reference-based metrics «спростовують самі себе» у QG, посилює цей висновок: єдине reference question не дає надійної оцінки якості [11, 21].

Тому в статті доцільно поєднувати кілька типів метрик. BERTScore і BLEURT корисні як semantic/reference-based показники; PMAN і RQUGE важливі тому, що зміщують акцент до answerability і можуть краще відображати придатність питання для реального використання; retrieval-style Hit@k залишається зручною sanity-check метрикою для попередньої перевірки семантичного зв'язку питання з твітом [12, 13, 14, 15].

Попри суттєвий прогрес у crisis informatics і наявність великих кризових корпусів, більшість наявних датасетів орієнтована на tweet-level classification, informativeness detection, humanitarian categorization, priority estimation або multimodal annotation. HumAID і CrisisBench надають основу насамперед для задач класифікації та порівняння моделей на рівні повідомлень; AIDR і TREC-IS

ISSN-print 1991-7848
ISSN-online 2707-9457

акцентують автоматичну фільтрацію та виокремлення actionable information; CrisisMMD розширює аналіз завдяки мультимодальності, але зберігає логіку анотаційних задач. Водночас ці ресурси практично не покривають сценарій, у якому користувацький запит залишається абстрактним, а коректна відповідь визначається прихованим кризовим контекстом і має бути відновлена моделлю без прямих маркерів типу лиха в самому питанні. Отже, для верифікації прихованої контекстної адаптації LLM необхідний спеціалізований датасет, що містить пари виду «абстрактний запит — кризово-залежна відповідь», де питання семантично пов'язане з кризовим повідомленням, але не підказує подію напряму [2, 3, 4, 5, 6].

Мета дослідження

Метою роботи є розроблення методу побудови кризово-контекстного датасету для подальшої верифікації Adaptive IRM у задачах прихованої контекстної адаптації відповідей великих мовних моделей. Для досягнення цієї мети необхідно: вибрати кризовий вихідний корпус; визначити цільові категорії подій; перетворити кризові повідомлення на пари «абстрактне питання — кризово-залежна відповідь»; вилучити з питань прямі кризові маркери; реалізувати автоматичний контроль формальних обмежень генерації; сформувати підсумковий датасет; підготувати план ручної та автоматичної перевірки якості корпусу; задати event-disjoint experimental split і визначити майбутній протокол порівняння LLM-baseline, LLM+Adaptive IRM і зіставних контекстних baselines. Така постановка узгоджується з наявною літературою щодо disaster LLMs, domain adaptation evaluation і question generation evaluation, але закриває окрему, досі недостатньо представлену задачу [7, 10, 11].

Викладення основного матеріалу дослідження

Вихідні дані та конструкція датасету

Як вихідна база використовується HumAID, оскільки цей корпус достатньо великий, розмічений за реальними disaster events і вже застосовувався як ресурс для кризового машинного навчання [2]. У поточній версії датасету сформовано 41 152 записи за п'ятьма категоріями: hurricanes — 27 443; earthquakes — 6 974; cyclones — 3 933; wildfires — 2 242; floods — 560.

На відміну від вихідної класифікаційної логіки кризових корпусів, у запропонованій постановці кожен приклад перетворюється на пару виду «абстрактний запит — кризово-залежна відповідь», де питання має бути загальним і не містити слів-маркерів лиха, а вихідний твіт виступає референсною кризово-залежною відповіддю. Така конструкція створює controlled setting для перевірки того, чи зможе окремий адаптаційний механізм повернути у відповідь пропущений контекст.

Метод генерації питань

Метод побудови датасету охоплює кілька послідовних етапів: відбір вихідних кризових повідомлень, генерацію абстрактного питання, автоматичну перевірку формальних обмежень, повторну генерацію в разі порушення умов і fallback-механізм для збереження повноти корпусу.

Для кожного повідомлення з HumAID локально розгорнута LLM генерувала коротке питання, на яке вихідний твіт має давати пряму відповідь. Як генеративна модель використовувалася openai/gpt-oss-20b у форматі GGUF з MXFP4-квантизацією. Інференс виконувався через локальний endpoint, що давало змогу проводити генерацію без звернення до зовнішніх API й фіксувати використовувану конфігурацію на рівні експерименту.

Під час генерації використовувалися стохастичні параметри декодування: temperature = 0.8, top-k = 40, top-p = 0.8, min-p = 0.05, repetition penalty = 1.1. Рівень reasoning effort було встановлено як Low. Фіксований random seed не задавався, тому генерація є стохастичною, а повне посимвольне відтворення окремих питань під час повторного запуску не гарантується. Водночас відтворюваними залишаються сама процедура побудови корпусу, набір обмежень, prompt-flow і правила автоматичної валідації.

Метою генерації було отримати не довільне питання до твіта, а питання спеціального типу: воно має бути достатньо конкретним, щоб вихідний твіт був його прямою відповіддю, але водночас не має містити прямої вказівки на тип лиха. Тому prompt було побудовано навколо трьох ключових вимог:

1. питання має відображати головний зміст твіта;
2. питання має бути WH-питанням;
3. питання не має містити заборонених кризових маркерів.

ISSN-print 1991-7848
ISSN-online 2707-9457

У primary generation prompt моделі задавалася інструкція написати рівно одне питання. В інструкції зазначалося, що твіт має бути найкращою прямою відповіддю на це питання, а саме питання має бути коротким, природним і пов'язаним з одним конкретним фактом: числом, людиною, організацією, дією, допомогою, місцем, часом, пошкодженням, підтримкою або оновленням. Також явно заборонялися yes/no-питання, питання про хештеги, авторів, сам твіт, а також близьке копіювання вихідного формулювання. Приклад загальної логіки primary prompt наведено на Лістингу 1.

```
Task: write exactly ONE question.

Goal:
The tweet text must be the best direct answer.

Rules:
- Ask about the MAIN point, not a side detail.
- Ask only for information stated in the tweet.
- Use one WH-question only.
- No yes/no questions.
- Do NOT mention any disaster, crisis, emergency, hazard, or
event.
- Do NOT copy or closely paraphrase the tweet.
- Keep it short.
- End with "?".
```

Лістинг 1 – Приклад primary generation prompt для формування одного підсумкового питання за текстом твіта

Користувацький prompt для primary generation мав мінімальну структуру: моделі передавався текст твіта, після чого вимагалося повернути лише одне підсумкове питання. Приклад користувацького prompt наведено на Лістингу 2.

```
Tweet text:
<tweet>
{{TWEET_TEXT}}
</tweet>

Write one natural question that this tweet answers directly.
Return only the question.
```

Лістинг 2 – Структура користувацького prompt для передавання тексту твіта та отримання одного підсумкового питання

Якщо згенероване питання не проходило автоматичну перевірку, запускалася retry generation. У retry prompt додатково передавалося попереднє невалідне питання, а модель отримувала інструкцію не повторювати його формулювання й згенерувати кращу заміну. Retry prompt був жорсткішим: вимагав коротшого WH-питання, фокусу на головному повідомленні твіта й заміни потенційних кризових слів нейтральними родовими іменниками, такими як amount, support, damage, group, help, place або update.

Загальна структура retry user prompt наведена на Лістингу 3.

```
    Tweet text:
    <tweet>
    {{TWEET_TEXT}}
    </tweet>

Bad previous question:
    <bad_question>
{{BAD_QUESTION}}
</bad_question>

Do not reuse its wording.
    Write the single best replacement question.
    Return only the question.
```

Лістинг 3 – Структура retry user prompt для повторної генерації питання після невдалої автоматичної перевірки

Таким чином, генерація будувалася не як одноразовий виклик LLM, а як контрольований pipeline: спочатку виконувалася основна генерація, потім

ISSN-print 1991-7848
ISSN-online 2707-9457

формальна перевірка, далі повторна генерація за потреби, і лише після цього fallback.

Заборонені кризові маркери

Для запобігання прямій підказці типу лиха використовувався фіксований список заборонених слів. Він охоплював прямі назви природних і техногенних подій, а також загальні crisis-related терміни:

disaster, earthquake, quake, aftershock, aftershocks, tremor, tsunami, hurricane, cyclone, typhoon, tornado, storm, blizzard, wildfire, fire, flood, landslide, mudslide, drought, eruption, volcano, pandemic, outbreak, explosion, blast, evacuation, catastrophe, emergency.

Цей список використовувався одночасно на двох рівнях. По-перше, він включався в prompt як явне обмеження для LLM. По-друге, той самий список застосовувався в автоматичному валідаторі після генерації. Таке дублювання було необхідним, оскільки prompt-інструкції не гарантують повного дотримання обмежень, особливо за стохастичного декодування.

Автоматична валідація питань

Після генерації кожне питання проходило автоматичну перевірку. Питання вважалося валідним лише за виконання кількох умов. Воно не має бути порожнім, має починатися з WH-слова, закінчуватися знаком питання, не містити заборонених кризових маркерів, не бути yes/no-питанням і не перевищувати встановлене обмеження довжини. Додатково відкидалися шаблонні або надто слабкі формулювання, наприклад питання виду What is the key detail about…, What is reported about…, What is mentioned about…, What is the situation… і близькі до них конструкції.

На рівні валідатора перевірялися такі групи обмежень:

- наявність тексту питання;
- відсутність слів із forbidden list;
- початок питання з WH-маркера: what, who, when, where, why, which, how;
- відсутність yes/no-структури на початку питання;
- наявність фінального знака ?;
- відсутність наперед заданих bad patterns;
- обмеження довжини питання.

Якщо питання порушувало хоча б одну з цих умов, воно не включалося до підсумкового датасету напряму. Натомість виконувалася retry generation. Якщо повторна генерація також не давала валідного результату, застосовувався fallback-механізм. Fallback використовувався як допоміжна процедура, що дає змогу зберегти запис у корпусі навіть тоді, коли LLM не змогла згенерувати питання, яке задовольняє всі обмеження.

У поточній версії корпусу fallback було використано 1 432 рази, що становить 3.48% від загальної кількості записів. Це значення характеризує стійкість генеративного pipeline: для більшої частини корпусу питання було отримано на етапі primary або retry generation, а fallback застосовувався лише в обмеженій частці випадків. Водночас якість primary-, retry- і fallback-прикладів має бути окремо перевірена на наступному етапі за допомогою формалізованої ручної та автоматичної валідації.

Підсумкове формулювання методу

Запропонований метод будує пари виду «абстрактний запит — кризово-залежна відповідь» на основі кризових повідомлень HumAID [2]. Для кожного повідомлення локально розгорнута LLM openai/gpt-oss-20b генерує коротке WH-питання, яке має відображати основний зміст вихідного твіта, але не містити прямих кризових маркерів. Генерація виконується у стохастичному режимі з фіксованими параметрами декодування, після чого кожне питання проходить автоматичну перевірку за формальними критеріями: наявність WH-структури, відсутність yes/no-форми, відсутність заборонених слів, коректне завершення знаком питання, обмеження довжини й фільтрація небажаних шаблонних формулювань.

У разі порушення цих вимог виконується повторна генерація з жорсткішою інструкцією та передаванням попереднього невалідного питання як негативного прикладу. Якщо повторна генерація також не приводить до валідного результату, використовується fallback-питання. У результаті формується кризово-контекстний датасет, у якому користувацький запит навмисно залишається абстрактним, а правильна інтерпретація відповіді потребує врахування прихованого контексту події. Така постановка відрізняється від класичних задач кризової класифікації та призначена для подальшої верифікації прихованої контекстної адаптації LLM [2, 3, 7].

ISSN-print 1991-7848
ISSN-online 2707-9457

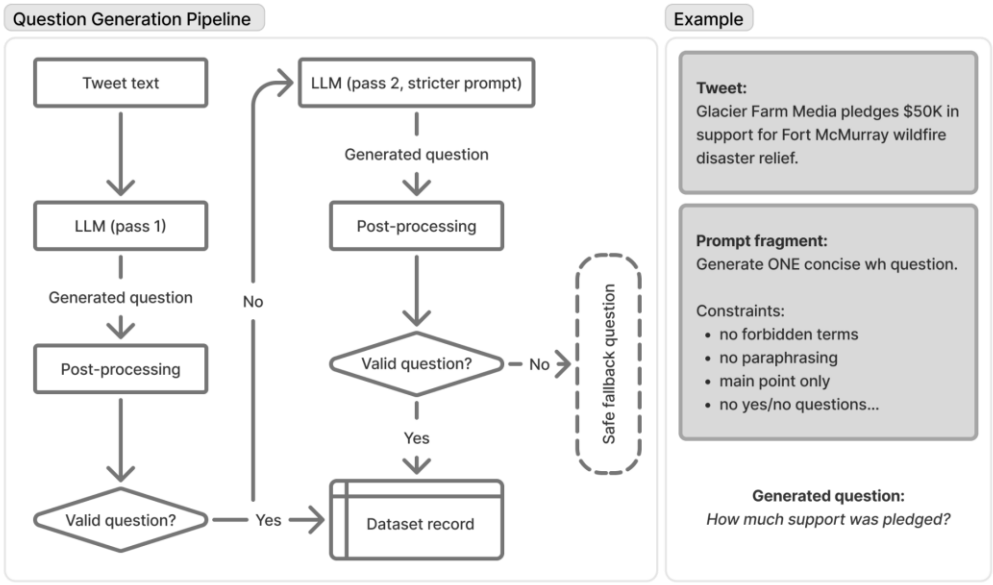

Рисунок 1. Конвеєр генерації запитань для побудови кризо-незалежних запитів на основі твітів.

Для заданого твіту LLM генерує кандидатне запитання (перший прохід), яке далі проходить постобробку та перевірку. Якщо перевірка не пройдена, виконується другий прохід генерації з використанням суворішого промпта. Якщо обидві спроби є неуспішними, використовується безпечне запасне запитання. Запитання вважається коректним, якщо воно не містить заборонених термінів, пов’язаних із кризами, не є перефразуванням вхідного твіту, відповідає формату WH-запитання та задовольняє обмеження довжини. Постобробка включає нормалізацію, зокрема виділення одного запитання та забезпечення коректного форматування. Фрагмент промпта наведено у скороченому вигляді з ілюстративною метою.

Результати

У результаті застосування запропонованого методу було сформовано кризово-контекстний датасет обсягом 41 152 записи. Датасет охоплює п'ять категорій подій: hurricanes, earthquakes, cyclones, wildfires і floods. Розподіл записів за категоріями є нерівномірним: найбільшу частку становлять повідомлення, пов'язані з ураганами, тоді як категорія повеней представлена суттєво меншою кількістю прикладів.

Таблиця 1

Розподіл записів сформованого кризово-контекстного датасету за категоріями подій

Категорія	Кількість записів
hurricanes	27 443
earthquakes	6 974
cyclones	3 933
wildfires	2 242
floods	560
Усього	41 152

Для забезпечення стійкості пайплайна використовувалася схема primary generation -> retry generation -> fallback. Fallback було задіяно у 1 432 випадках, що становить 3.48% загального обсягу корпусу. Це значення показує, що в більшості випадків питання були сформовані основним або повторним проходом, а fallback використовувався як допоміжний механізм для збереження повноти датасету.

Таблиця 2

Узагальнені показники сформованого датасету та частка використання fallback-механізму

Показник	Значення
Загальна кількість записів	41 152
Кількість категорій	5
Fallback	1 432
Fallback, %	3.48

Основним результатом цього етапу є отримання відтворюваного методу генерації кризово-контекстних пар і сформованого на його основі датасету. Запропонований корпус фіксує постановку, у якій питання залишається абстрактним і не містить прямої вказівки на тип лиха, тоді як відповідь зберігає зв'язок із конкретним кризовим повідомленням.

На наступному етапі планується провести формалізовану перевірку якості датасету. Вона має включати ручну стратифіковану валідацію, аналіз primary-, retry- і fallback-прикладів, а також автоматичну retrieval-style перевірку семантичної узгодженості між згенерованими питаннями та вихідними кризовими повідомленнями. Тому будь-які метрики якості питань і їхнього

ISSN-print 1991-7848
ISSN-online 2707-9457

зв'язку з вихідними твітами розглядаються в цій статті як частина майбутнього протоколу валідації, а не як уже завершений експериментальний результат.

План валідації та експериментальний дизайн

План ручної валідації

Наступним обов'язковим кроком є формалізована ручна перевірка якості датасету. Для цього пропонується використовувати стратифіковану вибірку з 500 прикладів, по 100 на кожен клас. Такий дизайн є кращим порівняно з пропорційною вибіркою, оскільки дає змогу забезпечити достатнє представництво рідкісних категорій, насамперед floods, які за пропорційного відбору могли б бути представлені надто малою кількістю прикладів. Для підвищення надійності оцінювання перевірку доцільно проводити двома незалежними анотаторами з подальшим розрахунком міжанотаторної узгодженості.

Критерії ручного оцінювання мають бути задані заздалегідь і застосовуватися однаково до всіх прикладів аудиту. Для кожного прикладу необхідно перевірити, чи є питання валідним питальним реченням; чи зберігає воно семантичний зв'язок із вихідним твітом; чи можна відповісти на нього безпосередньо за текстом твіта; чи відсутні в ньому прямі кризові маркери; чи не є воно майже дослівною копією вихідного повідомлення; чи є його формулювання природним і зрозумілим.

Такий протокол узгоджується із сучасними підходами до оцінювання question generation, у яких якість питання визначається не лише формальною подібністю до референсу, а й його answerability, consistency, clarity і придатністю для відповіді за вихідним контекстом. Зокрема, QGEval акцентує багатовимірне оцінювання якості питань, тоді як PMAN і RQUGE додатково підкреслюють важливість перевірки того, чи можна отримати відповідь на згенероване питання із заданого контексту [11, 14, 15].

За підсумками ручного аудиту планується зафіксувати кілька груп показників: загальну частку валідних питань, частку валідних питань за кожним кризовим класом, частоту порушень, пов'язаних із наявністю прямих crisis markers, рівень міжанотаторної узгодженості, а також типологію виявлених помилок. Мінімальна типологія помилок охоплює надто загальне питання, втрату answerability, приховане відсилання до типу лиха, near-copy

вихідного твіта й неприродне формулювання. Такий аналіз дасть змогу використовувати ручну перевірку не лише як інструмент контролю якості, а й як джерело дослідницьких висновків про типові обмеження запропонованого методу генерації питань.

Експериментальний план наступного етапу

Наступним етапом дослідження є реалізація Adaptive IRM і проведення контрольованого порівняння моделей в event-disjoint постановці. Оскільки HumAID охоплює 19 окремих disaster events [2], train/val/test split доцільно формувати на рівні подій, щоб уникнути інформаційного витоку між вибірками. В експериментальній частині планується порівняти базову LLM без додаткової адаптації, модель з Adaptive IRM, retrieval-augmented baseline і parameter-efficient baseline на основі PEFT. Як метрики передбачається використовувати retrieval-style Hit@k, semantic metrics BERTScore і BLEURT, answerability-oriented metrics PMAN та/або RQUGE, а також формалізоване ручне оцінювання. Такий дизайн дасть змогу перевірити, чи справді Adaptive IRM покращує відновлення прихованого кризового контексту порівняно як із чистою LLM, так і з альтернативними механізмами контекстного підсилення [2, 8, 9, 10, 11].

Висновки

У роботі запропоновано метод побудови кризово-контекстного датасету для подальшої верифікації Adaptive IRM у задачах прихованої контекстної адаптації відповідей великих мовних моделей. На відміну від наявних crisis datasets, орієнтованих переважно на tweet-level classification, informativeness detection і humanitarian categorization [2, 3, 4, 5, 6], запропонований корпус моделює ситуацію, у якій користувацький запит залишається абстрактним, а коректна відповідь визначається прихованим кризовим контекстом.

Основним результатом дослідження є сформований датасет обсягом 41 152 записи за п'ятьма категоріями кризових подій: hurricanes, earthquakes, cyclones, wildfires і floods. Для побудови корпусу було запропоновано генеративну схему, що включає primary generation, retry generation і fallback-механізм. Fallback було використано у 1 432 випадках, що становить 3.48% корпусу й характеризує роботу пайплайна з погляду повноти формування датасету.

ISSN-print 1991-7848
ISSN-online 2707-9457

Таким чином, внесок роботи полягає в розробленні методу перетворення кризових повідомлень на пари виду «абстрактний запит — кризово-залежна відповідь» і в отриманні датасету, придатного для подальшої перевірки прихованої контекстної адаптації LLM. Валідація якості згенерованих питань, event-disjoint split, реалізація Adaptive IRM і порівняння з baseline-підходами розглядаються як задачі наступного етапу дослідження.

У подальшій роботі необхідно: провести стратифіковану ручну валідацію корпусу на 500 прикладах із двома анотаторами; окремо проаналізувати якість primary-, retry- і fallback-прикладів; виконати автоматичну retrieval-style перевірку семантичної узгодженості питань і вихідних кризових повідомлень; зафіксувати event-disjoint split; реалізувати Adaptive IRM; виконати експериментальне порівняння LLM baseline, LLM+Adaptive IRM, RAG і PEFT-baseline; розширити набір метрик за межі формального контролю генерації. [7, 8, 9, 10, 11].

ЛІТЕРАТУРА/ REFERENCES

1. Reuter C., Hughes A. L., Kaufhold M.-A. Social media in crisis management: An evaluation and analysis of crisis informatics research. International Journal of Human–Computer Interaction. 2018. Vol. 34, No. 4. P. 280–294. DOI: 10.1080/10447318.2018.1427832.
2. Alam F., Qazi U., Imran M., Ofli F. HumAID: Human-Annotated Disaster Incidents Data from Twitter with Deep Learning Benchmarks. Proceedings of the International AAAI Conference on Web and Social Media. 2021. Vol. 15, No. 1. P. 933–942. DOI: 10.1609/icwsm.v15i1.18116.
3. Alam F., Sajjad H., Imran M., Ofli F. CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing. Proceedings of the International AAAI Conference on Web and Social Media. 2021. Vol. 15, No. 1. P. 923–932. DOI: 10.1609/icwsm.v15i1.18115.
4. Imran M., Castillo C., Lucas J., Meier P., Vieweg S. AIDR: Artificial Intelligence for Disaster Response. WWW '14 Companion: Proceedings of the 23rd International Conference on World Wide Web. New York : ACM, 2014. P. 159–162. DOI: 10.1145/2567948.2577034.
5. McCreadie R., Buntain C., Soboroff I. TREC Incident Streams: Finding Actionable Information on Social Media. Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2019). Valencia, Spain : ISCRAM Association, 2019. P. 691–705.
6. Alam F., Ofli F., Imran M. CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. Proceedings of the International AAAI Conference on Web and Social Media. 2018. Vol. 12, No. 1. DOI: 10.1609/icwsm.v12i1.14983.
7. Lei Z., Dong Y., Li W., Ding R., Wang Q. R., Li J. Harnessing Large Language Models for Disaster Management: A Survey. Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria : Association for Computational Linguistics, 2025. P. 14528–14551. DOI: 10.18653/v1/2025.findings-acl.750.

8. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W.-t., Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 9459–9474. DOI: 10.5555/3495724.3496517.
9. Han Z., Gao C., Liu J., Zhang J., Zhang S. Q. Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. Transactions on Machine Learning Research. 2024. URL: https://openreview.net/forum?id=lIsCS8b6zj (дата звернення: 29.04.2026).
10. Afzal A., Chalumattu R., Matthes F., Mascarell L. AdaptEval: Evaluating Large Language Models on Domain Adaptation for Text Summarization. Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U). Miami, Florida, USA : Association for Computational Linguistics, 2024. P. 76–85. DOI: 10.18653/v1/2024.customnlp4u-1.8.
11. Fu W., Wei B., Hu J., Cai Z., Liu J. QGEval: Benchmarking Multi-dimensional Evaluation for Question Generation. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA : Association for Computational Linguistics, 2024. P. 11783–11803. DOI: 10.18653/v1/2024.emnlp-main.658.
12. Zhang T., Kishore V., Wu F., Weinberger K. Q., Artzi Y. BERTScore: Evaluating Text Generation with BERT. arXiv preprint arXiv:1904.09675. 2019. DOI: 10.48550/arXiv.1904.09675.
13. Sellam T., Das D., Parikh A. P. BLEURT: Learning Robust Metrics for Text Generation. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online : Association for Computational Linguistics, 2020. P. 7881–7892. DOI: 10.18653/v1/2020.acl-main.704.
14. Wang Z., Funakoshi K., Okumura M. Automatic Answerability Evaluation for Question Generation. arXiv preprint arXiv:2309.12546. 2023. DOI: 10.48550/arXiv.2309.12546.
15. Mohammadshahi A., Scialom T., Yazdani M., Yanki P., Fan A., Henderson J., Saeidi M. RQUGE: Reference-Free Metric for Evaluating Question Generation by Answering the Question. Findings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada : Association for Computational Linguistics, 2023. P. 6845–6867. DOI: 10.18653/v1/2023.findings-acl.428.
16. Houlsby N., Giurgiu A., Jastrzebski S., Morrone B., de Laroussilhe Q., Gesmundo A., Attariyan M., Gelly S. Parameter-Efficient Transfer Learning for NLP. Proceedings of the 36th International Conference on Machine Learning. 2019. Vol. 97. P. 2790–2799. URL: https://proceedings.mlr.press/v97/houlsby19a.html (дата звернення: 29.04.2026).
17. Li X. L., Liang P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Online : Association for Computational Linguistics, 2021. P. 4582–4597. DOI: 10.18653/v1/2021.acl-long.353.
18. Hu E. J., Shen Y., Wallis P., Allen-Zhu Z., Li Y., Wang S., Wang L., Chen W. LoRA: Low-Rank Adaptation of Large Language Models. Proceedings of the International Conference on Learning Representations. 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9 (дата звернення: 29.04.2026).
19. Liu H., Tam D., Muqeeth M., Mohta J., Huang T., Bansal M., Raffel C. Few-Shot Parameter-Efficient Fine-Tuning Is Better and Cheaper than In-Context Learning. Advances in Neural Information Processing Systems. 2022. Vol. 35. P. 1950–1965. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/0cde695b83bd186c1fd456302888454c-Abstract-Conference.html (дата звернення: 29.04.2026).

ISSN-print 1991-7848
ISSN-online 2707-9457

20. Guo S., Liao L., Li C., Chua T.-S. A Survey on Neural Question Generation: Methods, Applications, and Prospects. Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence. 2024. P. 8038–8047. DOI: 10.24963/ijcai.2024/889.
21. Nguyen B., Yu M., Huang Y., Jiang M. Reference-based Metrics Disprove Themselves in Question Generation. Findings of the Association for Computational Linguistics: EMNLP 2024. Miami, Florida, USA : Association for Computational Linguistics, 2024. P. 13651–13666. DOI: 10.18653/v1/2024.findings-emnlp.798.

Received 17.03.2026.
Accepted 15.04.2026.
Published 30.04.2026

UDC 004.8

Anton Guda, Mykyta Bereziuk

METHOD FOR CONSTRUCTING A CRISIS-CONTEXT DATASET FOR ADAPTIVE IRM VERIFICATION

Abstract. *Recent research in the field of crisis informatics is largely focused on the automatic processing of social media messages during emergency situations. Existing crisis corpora, including HumAID, CrisisBench, AIDR, TREC-IS, and CrisisMMD, provide an important foundation for message classification, informativeness detection, humanitarian categorization, prioritization, and multimodal annotation tasks. At the same time, most of these resources are oriented toward the analysis of individual messages or the identification of their class, rather than toward verifying the ability of a large language model to reconstruct hidden crisis context from an abstract query. With the development of large language models, there is a growing need for specialized datasets that make it possible to evaluate not only the general linguistic competence of a model, but also its ability to adapt a response to a context that is not explicitly specified in the user’s question.*

The purpose of this work is to develop a method for constructing a crisis-context dataset for the subsequent verification of Adaptive IRM in tasks of hidden contextual adaptation of large language model responses. To achieve this purpose, it is proposed to transform crisis messages from the HumAID corpus into pairs of the form “abstract query — crisis-dependent answer”, where the question does not contain direct markers of the disaster type but preserves a semantic connection with the original message.

The paper proposes a generative dataset construction pipeline that includes primary generation, retry generation, and a fallback mechanism. For each crisis message, a locally deployed large language model generates a short WH-question to which the original tweet should provide a direct answer. After generation, the question undergoes automatic validation according to formal criteria: the presence of an interrogative

structure, the absence of a yes/no form, ending with a question mark, compliance with the length limit, the absence of undesirable template-like formulations, and the absence of direct crisis markers such as earthquake, hurricane, flood, disaster, emergency, and others. If the initial question does not meet the specified requirements, a retry generation step is performed using a stricter instruction. In case of a repeated failure, a fallback question is applied, which makes it possible to preserve the completeness of the corpus. As a result, a dataset of 41,152 records was formed across five categories of crisis events: hurricanes, earthquakes, cyclones, wildfires, and floods. The fallback mechanism was used in 1,432 cases, which accounts for 3.48% of the corpus.

The main result of the study is a method for transforming crisis messages into “abstract query — crisis-dependent answer” pairs and the dataset constructed on its basis for the future verification of hidden contextual adaptation in LLMs. The proposed approach differs from classical crisis datasets in that it models a situation in which the question does not directly indicate the type of disaster, while the correct answer requires taking into account the hidden context of the event. Future work includes formalized manual validation of the corpus, automatic retrieval-style verification of semantic consistency, construction of an event-disjoint split, implementation of Adaptive IRM, and comparison with LLM-baseline, RAG, and PEFT-baseline approaches.

Keywords: *crisis-context dataset, large language models, hidden contextual adaptation, Adaptive IRM, question generation, crisis informatics, HumAID, question generation, dataset validation.*

Гуда Антон Ігоревич – д.т.н., професор, каф. «Комп’ютерні інформаційні технології», Український державний університет науки і технологій, УДУНТ, ORCID 0000-0003-1139-1580, atu.guda@gmail.com.

Березюк Микита Олександрович – аспірант, каф. «Комп’ютерні інформаційні технології», Український державний університет науки і технологій, УДУНТ, ORCID 0009-0000-2205-3611, nikitber@hotmail.com.

Guda Anton – professor, dep. “Computer and Information Technology”, Ukrainian State University of Science and Technology, USUST, ORCID 0000-0003-1139-1580, atu.guda@gmail.com.

Bereziuk Mykyta – post-graduate student, dep. “Computer and Information Technology”, Ukrainian State University of Science and Technology, USUST, ORCID 0009-0000-2205-3611, nikitber@hotmail.com.

ISSN-print 1991-7848
ISSN-online 2707-9457