МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КРИВОРІЗЬКИЙ ДЕРЖАВНИЙ ПЕДАГОГІЧНИЙ УНІВЕРСИТЕТ
Фізико-математичний факультет
Кафедра інформатики та прикладної математики

«Допущено до захисту»
Завідувач кафедри
_____ Моісеєнко Н. В.
«____» _____ 2024 р.

Реєстраційний №____
«____» _____ 2024 р.

ЗАСТОСУВАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ ГЕНЕРАЦІЇ КОНТЕНТУ

Кваліфікаційна робота студента групи І-20
ступінь вищої освіти «бакалавр»
спеціальності 014.09 Середня освіта (Інформатика)
Слободянюка Артема Валерійовича

Керівник: доктор педагогічних наук, професор,
старший дослідник
Семеріков Сергій Олексійович

Оцінка:
Національна шкала _____
Шкала ECTS ____ Кількість балів ____
Голова ФК _____ _____
Члени ФК _____ _____
_____ _____
_____ _____
_____ _____

Кривий Ріг – 2024

ЗАПЕВНЕННЯ

Я, Слободянюк Артем Валерійович, розумію і підтримую політику Криворізького державного педагогічного університету з академічної доброчесності. Запевняю, що ця кваліфікаційна робота виконана самостійно, не містить академічного плагіату, фабрикації, фальсифікації. Я не надавав і не одержував недозволену допомогу під час підготовки цієї роботи. Використання ідей, результатів і текстів інших авторів мають покликання на відповідне джерело.

Із чинним Положенням про запобігання та виявлення академічного плагіату в роботах здобувачів вищої освіти Криворізького державного педагогічного університету ознайомлений. Чітко усвідомлюю, що в разі виявлення у кваліфікаційній роботі порушення академічної доброчесності робота не допускається до захисту або оцінюється незадовільно.

ЗМІСТ

ВСТУП

Обґрунтування вибору теми дослідження. Актуальність дослідження зумовлена стрімким розвитком та зростанням популярності методів машинного навчання для генерації текстового контенту у останні роки. Особливо значний інтерес до цієї галузі виник після представлення у 2022 році чат-боту ChatGPT від OpenAI [38], який надав користувачам інтерфейс на основі великої мовної моделі GPT для генерації текстів природною мовою. Поява таких систем відкриває нові можливості для автоматизації створення різноманітного текстового контенту, від звичайних повідомлень до технічної документації та творчих текстів. Водночас, стрімкий розвиток цієї сфери ставить нові виклики щодо оцінки якості, контролю та практичного застосування згенерованих текстів. Тому актуальним є дослідження сучасного стану розвитку методів машинного навчання для генерації текстового контенту, аналіз нових підходів, метрик оцінювання, наборів даних та сфер застосування, що з'явились у науковій літературі протягом останніх років (2022-2024). Такий огляд дозволить узагальнити та систематизувати актуальні здобутки у даній галузі, окреслити перспективні напрямки подальших досліджень та розробок.

Об'єкт дослідження – генерація текстового контенту.

Предмет дослідження – методи машинного навчання, метрики оцінювання ефективності моделей, набори даних, застосування та використані природні мови, що розглядаються в літературі 2022-2024 рр.

Мета дослідження – виконати систематичний огляд застосування штучних нейронних мереж для генерації текстового контенту (2022-2024).

Відповідно до мети визначено такі основні **завдання дослідження**:

1. Дослідити методи (підходи, архітектури) глибокого навчання для генерації тексту, які з'явились чи були зазначені у роботах 2022-2024 рр.

2. Розглянути метрики для оцінювання ефективності моделей генерації тексту, які з'явились чи були зазначені у роботах 2022-2024 рр.

3. Визначити набори даних для генерації тексту, описані у роботах 2022-2024 рр.

4. Дослідити нові застосування генерації тексту, описані у роботах 2022-2024 рр.

5. Визначити, які природні мови використовувались для генерації тексту у роботах 2022-2024 рр.

Систематичний аналіз літератури є основним **методом даного дослідження**, який дозволяє узагальнити та синтезувати інформацію з великої кількості наукових публікацій (вторинних джерел) за чітко визначеною методикою. Для проведення огляду було обрано методику PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), яка є загальновизнаним стандартом для систематичних оглядів та мета-аналізів у різних галузях науки [46]. Систематичний аналіз за методикою PRISMA передбачає чітке планування дослідження, визначення критеріїв відбору публікацій, проведення ретельного пошуку літератури у провідних наукових базах даних, відбір релевантних досліджень, видобування та синтез даних. Такий підхід забезпечує повноту, надійність та відтворюваність отриманих результатів. Обраний метод повністю відповідає меті та завданням дослідження, дозволяючи отримати узагальнену картину сучасного стану досліджень у галузі генерації текстового контенту на основі аналізу значного масиву наукових публікацій за останні роки.

Новизна результатів дослідження полягає в тому, що вперше:

- узагальнено та систематизовано інформацію щодо передових методів глибокого навчання для генерації текстового контенту на основі аналізу літератури 2022-2024 рр.;

- проаналізовано нові метрики оцінювання ефективності моделей генерації тексту, які з'явились у дослідженнях останніх років;

- визначено основні набори даних для генерації текстового контенту, що використовувались у роботах 2022-2024 рр., та охарактеризовано їх особливості;

- виявлено нові сфери та задачі застосування моделей генерації текстів, які набули актуальності у дослідженнях 2022-2024 рр.;

- проаналізовано, які природні мови найчастіше використовувались для генерації текстового контенту у роботах останніх років.

Практичне значення результатів дослідження:

1. Результати огляду можуть бути використані дослідниками та розробниками для вибору актуальних методів, метрик та наборів даних при розробці систем генерації текстового контенту.

2. Інформація щодо нових сфер та задач застосування генерації текстів може допомогти у пошуку практичних задач, де дані методи будуть найбільш ефективними та затребуваними.

3. Огляд використання різних природних мов у дослідженнях генерації текстів має значення для розвитку мультилінгвальних систем та розширення можливостей автоматичного створення контенту різними мовами світу.

Структура та обсяг роботи. Робота складається зі вступу, трьох розділів, висновків до них, загальних висновків, списку використаних джерел (52 найменування), 1 додатку. Робота містить 12 таблиць та 10 рисунків. Загальний обсяг роботи – 61 сторінка.

РОЗДІЛ 1
ПОСТАНОВКА ПРОБЛЕМИ

1.1. Обґрунтування

Опрацювання природної мови (NLP – Natural Language Processing) – междисциплінарна галузь інформатики та лінгвістики [24, с. 1], класифікацію основних задач якої подано на рис. 1.1.

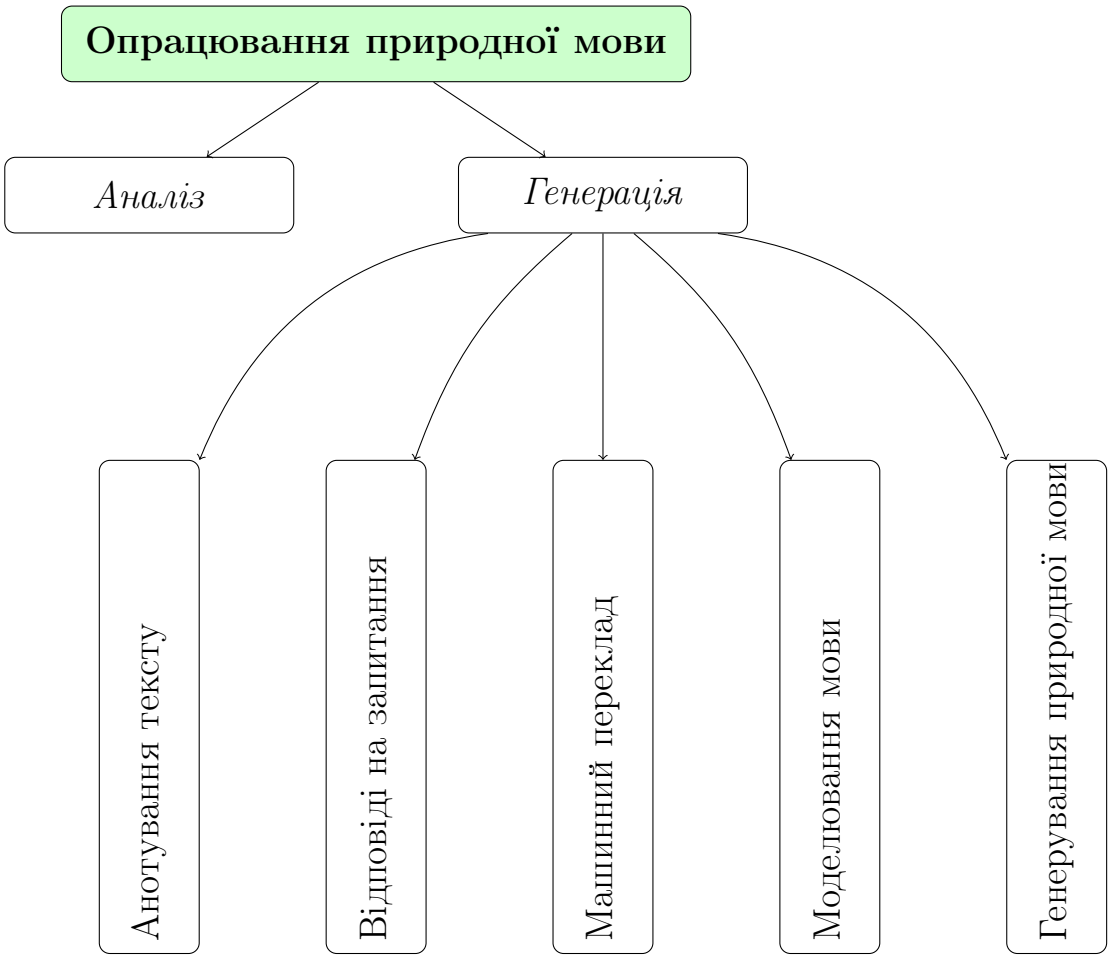

Рис. 1.1. Таксономія популярних задач NLP для генерації текста (на основі [24, с. 4]).

Генерація текстового контенту – розділ NLP, що поєднує обчислювальну лінгвістику та штучний інтелект для генерації тексту [5, с. 53490].

У 2022 році OpenAI [38] представив ChatGPT – чат-бот на основі моделі GPT, що надала користувачу інтерфейс природною мовою. В більшості систематичних оглядах розглядаються схожі питання, що пояснює вибір нашого огляду.

Попередній огляд “A Systematic Literature Review on Text Generation Using Deep Neural Network Models” [5] охоплював 90 джерел з 2015 по 2021 рр. Поява доступа до великих мовних моделей у 2022-2023 рр. [33] призвело до зростання інтересу до них (рис. 1.2), тому виникла потреба у доповненні попереднього огляду, основним результатом якого є класифікація (рис. 1.3):

1) *за архітектурою нейронної мережі*:

   - традиційні:
     - RNN – рекурентна нейронна мережа, що використовується для послідовних даних;
     - LSTM – мережа з довгою короткочасною пам’яттю, що працює краще за RNN за більших об’ємів даних;
     - GRU – вентильний рекурентний вузол (спрощена версія LSTM);
     - CNN – згорткова нейронна мережа.
   - інноваційні:
     - Attention Based – мережі, що використовують механізм уваги для підвищення значущості вхідних даних;
     - Transformer – мережі, що використовують механізм уваги без рекурентних або згорткових шарів;
     - BERT – розроблена Google нейронна мережа, що поєднує у собі механізми уваги без рекурентних або згорткових шарів із двонапрямленими кодувальниками

2) *за метриками якості*:

   - людино-орієнтовані:
     - Domain-Expert – залучення людини, яка є спеціалістом у даній галузі, для валідації результатів.
   - машинно-орієнтовані (автоматичні):
     - BLEU (bilingual evaluation understudy) – порівнює кількість і значення токенів (лексем) машинного і людського перекладу; значення слів до уваги не береться;

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation) – порівнює анотації/переклади, згенеровані машиною та людиною;
- Cosine Similartity – порівняння косинусів кута двох ненульових векторів: значення +1 відповідає однонапрямленим пропорційним векторам, -1 відповідає протилежно напрямленим пропорційним векторам;
- Content Selection – схожа до ROUGE метрика, яка використовує механізм уваги щодо заданої задачі;
- Diversity Score – метрика оцінки різноманітності.

3) *за застосуванням нейронної мережі*:

- AMR (Abstract Meaning Representation) – видобування семантичних співвідношень із тексту;
- Language Generation – генерація тексту, подібного до людського;
- Speech-to-text – перетворення мовлення на текст;
- Script Generation – генерація сценаріїв на основі заданих слів;
- Machine Translation – генерація машинного перекладу тексту з однієї мови на іншу;
- Text Summarization – генерація анотації до заданого тексту;
- Image Captioning – генерація опису до наданого зображення;
- Shopping Guide – генерація рекламного опису до наданого зображення товару;
- Weather Forecast – генерація тексту прогнозу погоди.

4) *за мовою генерації*:

- гарно забезпечених ресурсами: англійська, кітайська;
- недостатньо забезпечених ресурсами: бенгальська, корейська, балійська, іспанська, хінді, словацька, македонська.

5) *за набором даних для навчання нейронної мережі*:

- за типом розмітки:

- Labeled (розмічені дані);
- Unlabeled (нерозмічені дані);

- за типом:
  - Sentence – речення;
  - Paragraph – абзац;
  - Question/answer – дані типу питання та відповідь;
  - Document – дані у вигляді документу.

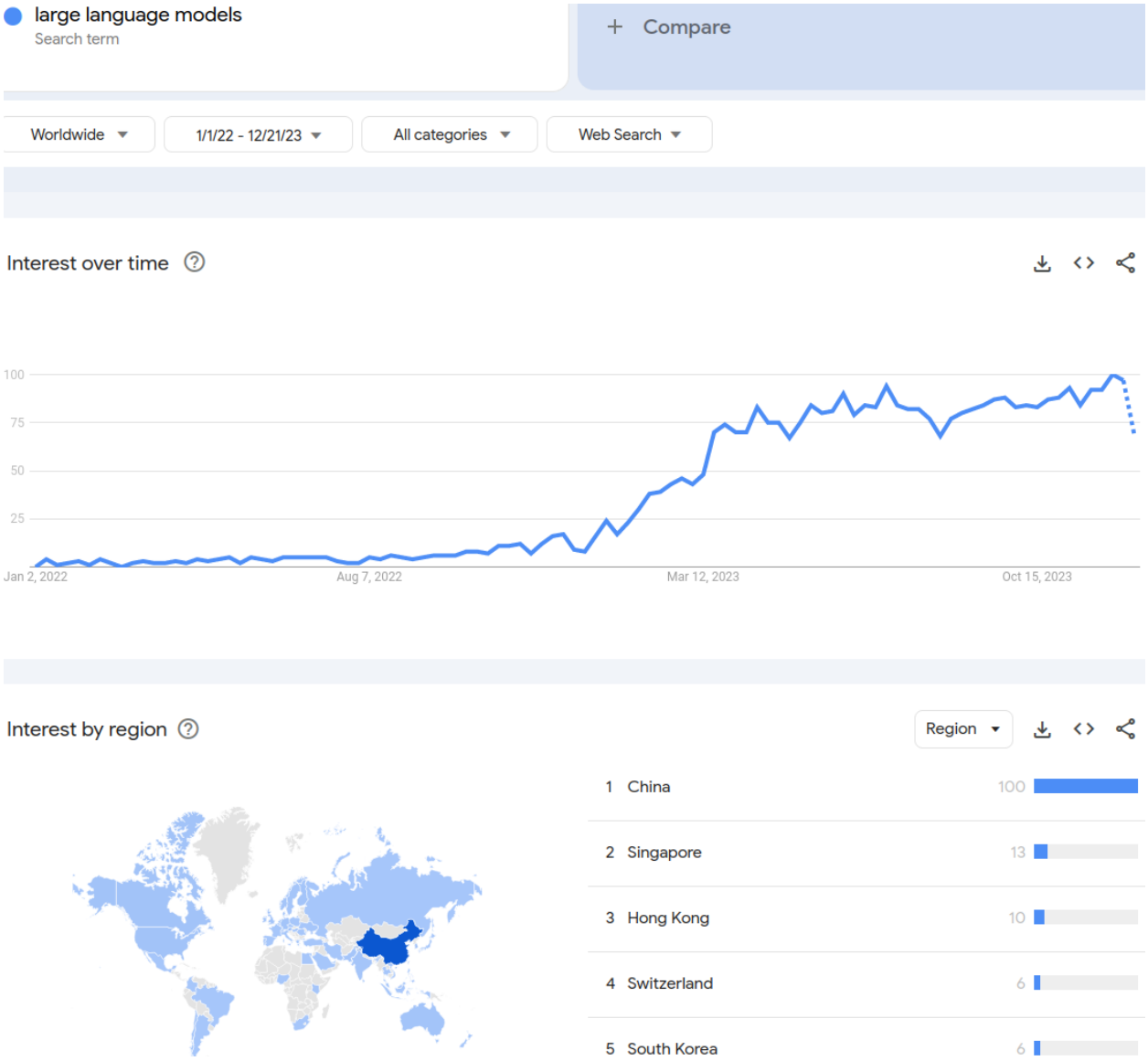

Рис. 1.2. Динаміка пошукових запитів за терміном “large language models” [33].

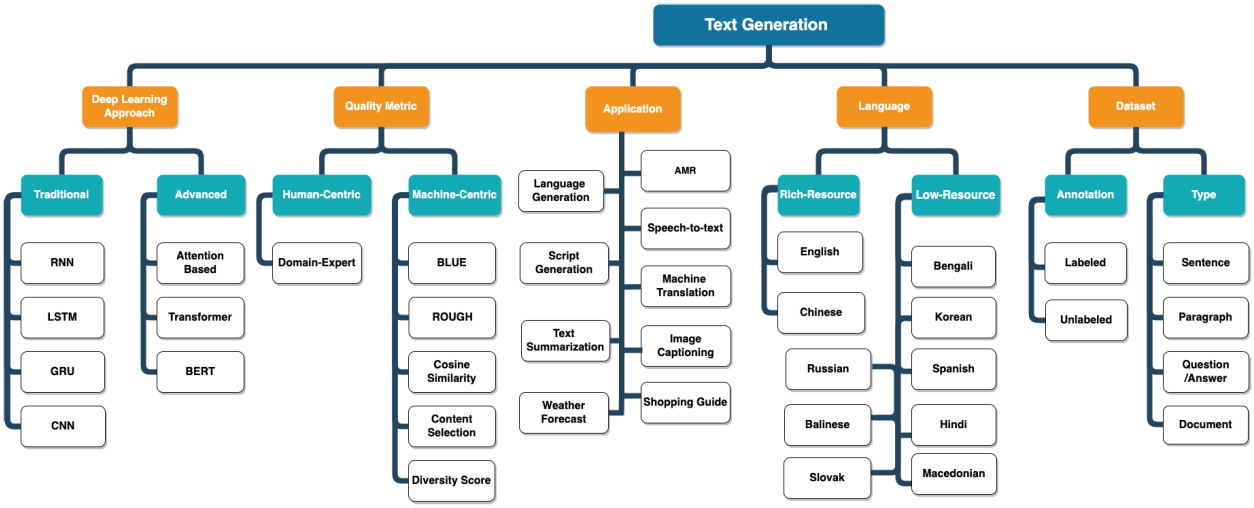

Рис. 1.3. Taxonomy of the text generation [5, с. 53493].

1.2. Дослідницькі задачі та питання

Для отримання результатів, представлених на рис. 1.3, Fatima та ін. [5] поставили таки задачі:

1. Дослідити існуючі традиційні та інноваційні методи (підходи) глибокого навчання для генерації тексту.
2. Розглянути метрики продуктивності для оцінювання моделей генерації тексту.
3. Дослідити методи оцінювання для вимірювання якості згенерованого тексту.
4. Оглянути нові галузі застосування генерування текстового контенту.
5. Обговорити найважливіші проблеми та майбутні напрями дослідження у галузі генерування текстового контенту.

Для доповнення отриманих Fatima та ін. [5] результатів дані *задачі дослідження* були уточнені:

1. Дослідити методи (підходи, архітектури) глибокого навчання для генерації тексту, які з'явились чи були зазначені у роботах 2022-2024 рр.
2. Розглянути метрики для оцінювання ефективності моделей генерації тексту, які з'явились чи були зазначені у роботах 2022-2024 рр.

3. Визначити набори даних для генерації тексту, описані у роботах 2022-2024 рр.

4. Дослідити нові застосування генерації тексту, описані у роботах 2022-2024 рр.

5. Визначити, які природні мови використовувались для генерації тексту у роботах 2022-2024 рр.

Аналогічно були уточнені *дослідницькі питання*:

ДП1. Які передові методи глибокого навчання використовуються для генерації тексту в літературі 2022-2024 рр.?

ДП2. Які нові метрики для оцінювання ефективності моделей генерації тексту в літературі 2022-2024 рр.?

ДП3. Які набори даних для генерації тексту описані в літературі 2022-2024 рр.?

ДП4. Які нові застосування генерації тексту описані в літературі 2022-2024 рр.?

ДП5. Які природні мови використовуються для генерації тексту в літературі 2022-2024 рр.?

Висновки до 1 розділу

У першому розділі представлено постановку проблеми дослідження генерації текстового контенту за допомогою штучних нейронних мереж у 2022-2024 рр. Зокрема:

1. Обґрунтовано необхідність оновлення попереднього систематичного огляду [5] через появу у 2022 році чат-ботів на основі великих мовних моделей, що спричинило зростання інтересу до генерації текстового контенту.

2. Наведено класифікацію методів генерації природної мови, отриману у попередньому огляді, за архітектурою нейронної мережі, метриками

якості, застосуванням, мовою генерації та типом набору даних для навчання.

3. Сформульовано уточнені задачі дослідження та дослідницькі питання щодо методів і метрик генерації та оцінювання текстового контенту, наборів даних, мов та застосувань, описаних у літературі 2022-2024 рр.

РОЗДІЛ 2
МЕТОДИКА

2.1. Джерела інформації та стратегія пошуку

Fatima та ін. [5] у попередньому огляді використали в якості надійних джерел даних 2 наукометричні бази даних (Web of Science та Scopus) та 4 бібліотеки (IEEE Xplore, SpringerLink, ScienceDirect та ACM Digital Library). Пошуковий запит по назвах статей, анотаціям та ключовим словам, використаний Fatima та ін. [5], поданий у табл. 2.1.

Таблиця 2.1

Групи обраних ключових слів [5, с. 53494].

Група 1: Слова, які відносяться до глибинного навчання	deep learning OR natural language processing OR NLP OR neural network OR RNN OR Recurrent OR Recursive OR LSTM OR GAN OR GPT-2 OR generative adversarial network
Група 2: Слова, які відносяться до генерації тексту	text generation OR language generation OR language modelling OR natural language generation OR neural language generation
Пошуковий запит	(Група 1) та (Група 2)

Наразі Scopus покриває близько 90% IEEE Xplore та ACM Digital Library, Web of Science – близько 50%; ScienceDirect та Scopus мають одного й того ж власника – Elsevier. Ураховуючи, що до Scopus входить значна частина вказаних бібліотек, замість 2 баз та 4 бібліотек була використана лише одна база – Scopus. Застосування пошукового запиту із попереднього огляду (табл. 2.1) дає 2580 документів за 2015–2020 рр. (проти 100 документів, вказаних у [5, с. 53494]). При пошуці виключно у назвах статей кількість документів зменшується до 109 і спостерігається часткове співпадення із переліком джерел [5, с. 53500—53503]).

Неможливість відтворення попередніх результатів за запитом із

табл. 2.1 спонукало до створення нового запиту:

```
(
    TITLE-ABS-KEY(neural network)
    OR
    TITLE-ABS-KEY(machine learning)
    OR
    TITLE-ABS-KEY(deep learning)
)
AND
TITLE("text generation")
```

Перша частина запиту була спрощена до трьох ключових фраз, дві з яких (“neural network” та “deep learning”) співпадають із першою групою табл. 2.1, а третя (“machine learning”) узагальнює усі інші ключові слова першої групи, включно з неіснуючими на момент створення попереднього огляду. До другої частини запиту була включена лише ключова фраза “text generation”, пошук якої виконується у заголовках документів (TITLE), а не в заголовках, анотаціях та авторських ключових словах (TITLE-ABS-KEY).

2.2. Критерії включення та виключення документів

Критерії включення:

1. Документи, опубліковані в період з 2022 по 2024 рік.
2. Документи, що стосуються генерації тексту за допомогою штучних нейронних мереж.
3. Документи, що описують підходи, архітектури, метрики якості, мови, набори даних або застосування генерації тексту.

Критерії виключення:

1. Документи, опубліковані до 2022 року або такі, що не містять даних за 2022-2024 рр.

2. Документи, які не стосуються генерації тексту або не використовують штучні нейронні мережі.

3. Документи, які не містять релевантної інформації щодо поставлених дослідницьких питань (нові методи, метрики, набори даних, застосування, природні мови).

2.3. Процес відбору документів

Запит до Scopus 04.03.2024 повернув 248 документів, розподіл яких за рокам подано на рис. 2.1. Із них 2 виявились дублікатами, а 157 – датованими раніше 2022 року, тому вони були виключені зі списку для отримання.

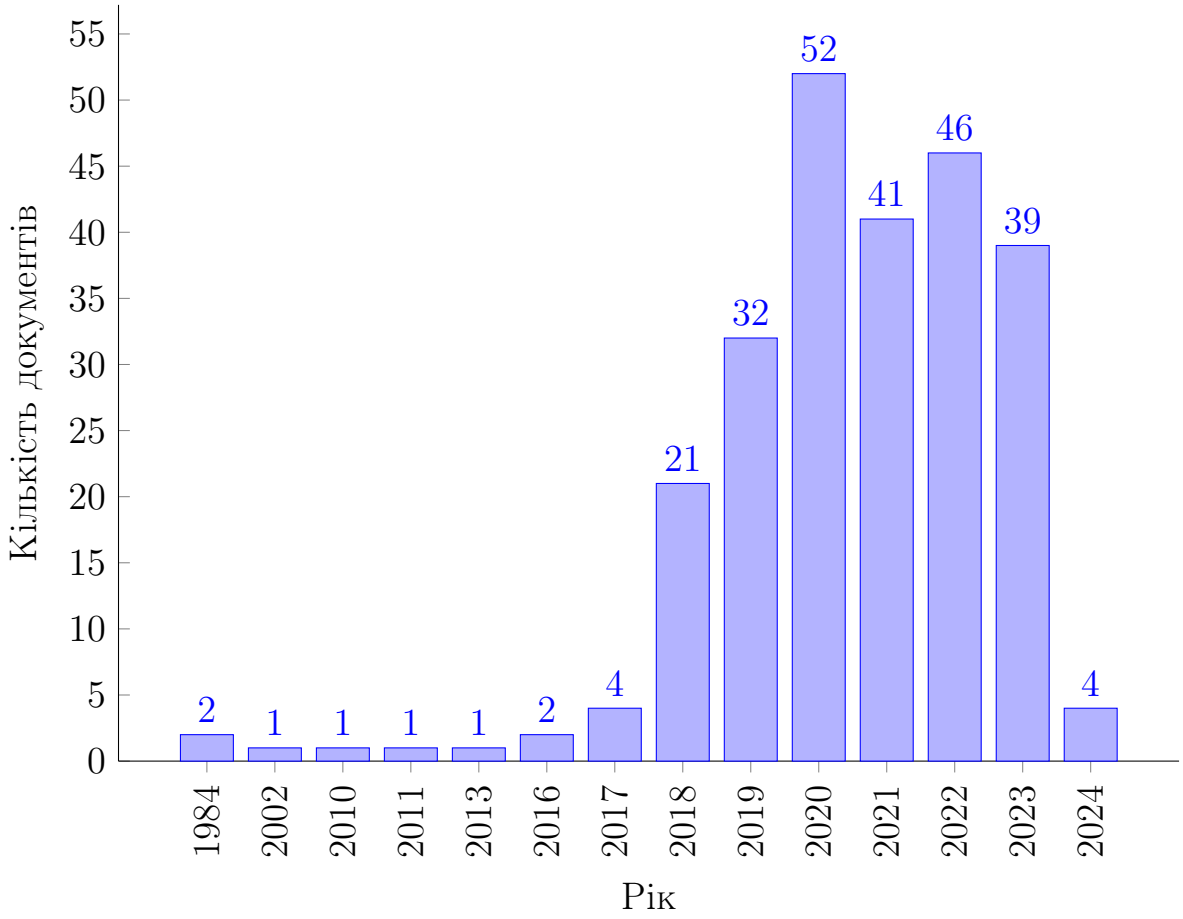

Рис. 2.1. Розподіл результатів пошуку за роками.

На рис. 2.2 подана схема відбору даних для систематичного огляду.

89 документів намагались отримати із сайтів видавців, із наукової соціальної мережі ResearchGate та серверів препринтів (насамперед arXiv). 41 документ (насамперед з ACM Digital Library та IEEE Xplore) отримати

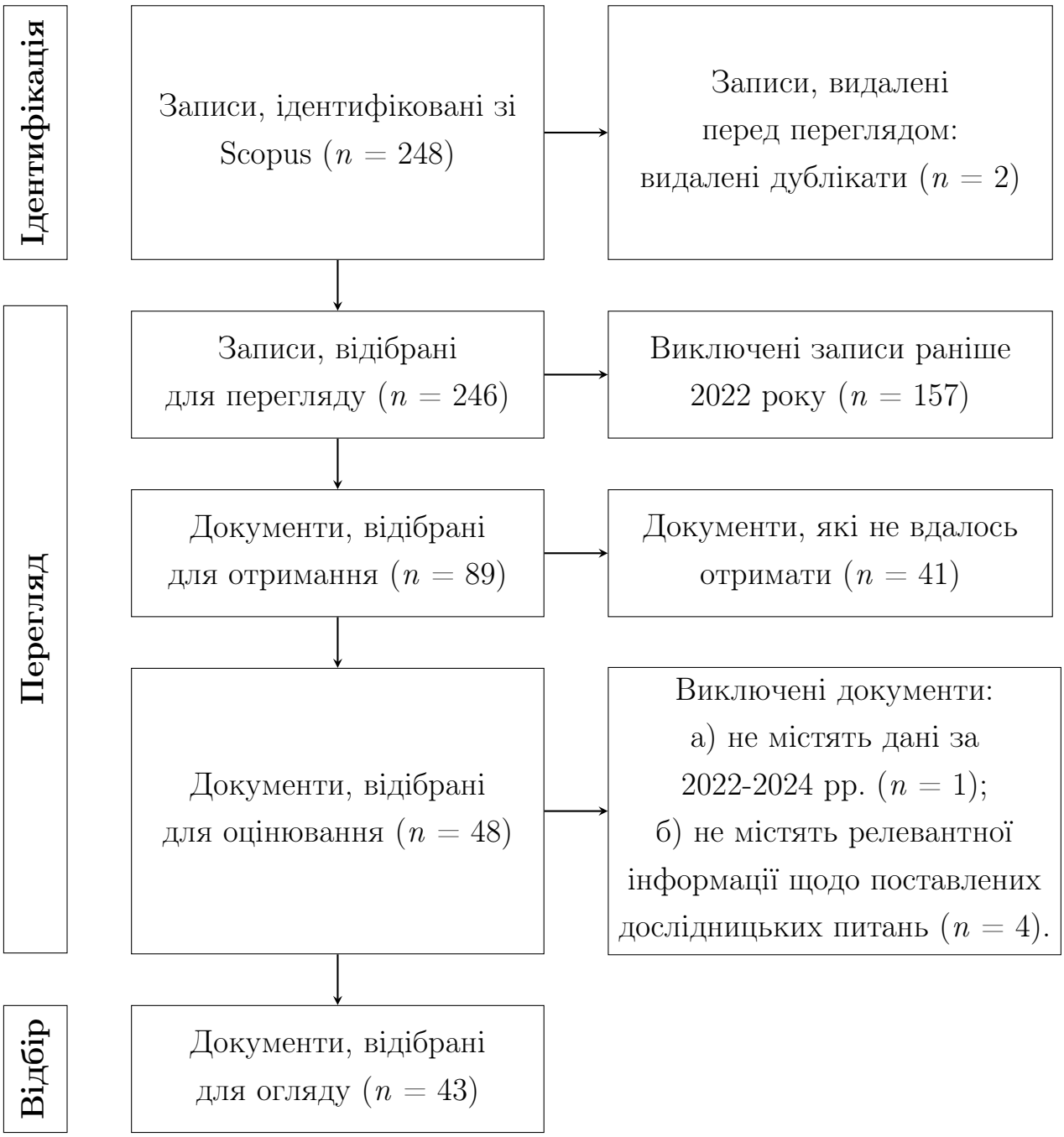

Рис. 2.2. Схема відбору даних для систематичного огляду (згідно методики PRISMA [46]).

не вдалось. Таким чином, для оцінювання було відібрано 48 документів, перегляд яких виявив 1 документ, що не містив дані за 2022-2024 рр., та 4 документи, що не містили релевантної інформації щодо поставлених дослідницьких питань.

43 документи були відібрані для огляду: [1—4; 6—9; 11—23; 25—27; 29; 31; 32; 34—37; 39—45; 47—51]. Огляд кожного документу виконувався згідно карти огляду (додаток 3.6). Для автоматизації видобування даних

за питаннями карти була застосована велика мовна модель (ВММ) Claude 3 Sonnet [30], на вхід якої подавався файл документу у форматі PDF із наступним промптом:

```
Опиши статтю за такими характеристиками:
Тип документу: стаття у журналі (ARTICLE) або стаття у
матеріалах конференції (CONFERENCE)
Назва
Рік публікації
Країни, які представляють автори
Мета статті
Використані архітектури нейронних мереж
Використані метрики якості
Характеристики використаних наборів даних - ім'я
Характеристики використаних наборів даних - тип даних: речення,
абзац, документ, питання-відповідь, не вказаний
Характеристики використаних наборів даних - розмір
Характеристики використаних наборів даних - формат: CSV, JSON,
XML, файли, не вказаний
Характеристики використаних наборів даних - за типом розмітки:
розмічені дані, нерозмічені дані
Характеристики використаних наборів даних - якість даних:
попередньо не опрацьовані (сирі), попередньо опрацьовані
Характеристики використаних наборів даних - за доступності:
загальнодоступний, приватний, не вказаний
Характеристики використаних наборів даних - посилання
Розв'язана задача генерації тексту (для чого застосовувалась
нейронна мережа)
Мова генерації тексту
```

Приклад відповіді подано на рис. 2.3.

Для зовнішньої перевірки була застосована ВВМ GPT 4 через розширення Scholar GPT [10]. У табл. 2.2 подано порівняння результатів автоматичного огляду статті [34].

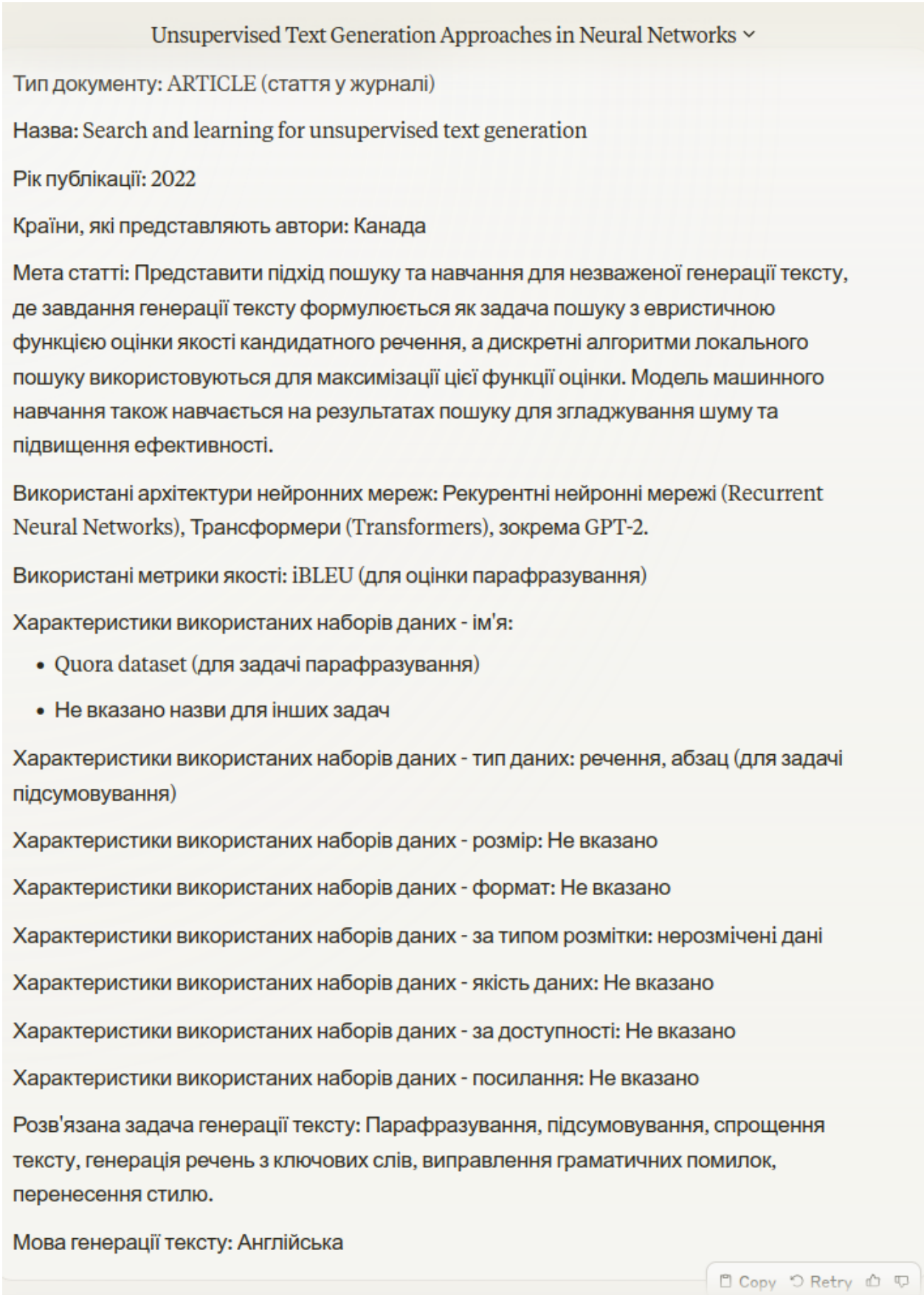

Рис. 2.3. Відповідь Claude 3 Sonnet на запит щодо складання карти огляду статті [34].

Таблиця 2.2

Порівняння результатів роботи великих мовних моделей Claude 3 Sonnet та GPT 4.

Запит	**Відповідь Claude 3 Sonnet**	**Відповідь GPT 4**
Тип документу: стаття у журналі (ARTICLE) або стаття у матеріалах конференції (CONFERENCE)	ARTICLE (стаття у журналі)	Article in a journal
Назва	Search and learning for unsupervised text generation	Search and Learning for Unsupervised Text Generation
Рік публікації	2022	2022
Країни, які представляють автори	Канада	Canada (The author, Lili Mou, is affiliated with the University of Alberta in Edmonton, Canada)
Мета статті	Представити підхід пошуку та навчання для незваженої генерації тексту, де завдання генерації тексту формулюється як задача пошуку з евристичною функцією оцінки якості кандидатного речення, а дискретні алгоритми локального пошуку використовуються для максимізації цієї функції оцінки. Модель машинного навчання також навчається на результатах пошуку для згладжування шуму та підвищення ефективності.	The article aims to introduce approaches to unsupervised text generation using search and learning methodologies. It focuses on overcoming the need for massive labeled datasets, which are required in supervised methods, thereby making text generation feasible for low-resource languages and reducing the human labor required for data annotation.
Використані архітектури нейронних мереж	Рекурентні нейронні мережі (Recurrent Neural Networks), Трансформери (Transformers), зокрема GPT-2	Neural architectures mentioned include recurrent neural networks (RNNs), Transformer architecture, and language models such as RoBERTa and pretrained language models by Radford et al., 2019.

Продовження на наступній сторінці

Табл. 2.2 – продовження із попередньої сторінки

Запит	**Відповідь Claude 3 Sonnet**	**Відповідь GPT 4**
Використані метрики якості	iBLEU (для оцінки парафразування)	Metrics such as language fluency, semantic coherency, and task-specific constraints (like sentence simplicity scores) are used to evaluate the quality of generated texts.
Характеристики використаних наборів даних		
- ім'я	Quora dataset (для задачі парафразування). Не вказано назви для інших задач	Not specified in detail for individual datasets.
- тип даних: речення, абзац, документ, питання-відповідь, не вказаний	речення, абзац (для задачі підсумовування)	Sentences, as the text generation involves tasks like paraphrasing, summarization, etc.
- розмір	Не вказано	Specific sizes are not mentioned, but the use of large-scale generic corpora is implied.
- формат: CSV, JSON, XML, файли, не вказаний	Не вказано	Not explicitly mentioned, but it includes labeled input-output pairs in various formats typical for NLP tasks.
- за типом розмітки: розмічені дані, нерозмічені дані	нерозмічені дані	Primarily deals with labeled data in supervised contexts, though the focus here is on reducing dependence on such datasets.
- якість даних: попередньо не опрацьовані (сирі), попередньо опрацьовані	Не вказано	The discussion includes the use of both raw and preprocessed data in different contexts.
- загальнодоступний, приватний, не вказаний	Не вказано	The text suggests a shift towards methods that do not require publicly available or parallel datasets.
- посилання	Не вказано	Not specific dataset URLs provided.

Продовження на наступній сторінці

Табл. 2.2 – продовження із попередньої сторінки

Запит	**Відповідь Claude 3 Sonnet**	**Відповідь GPT 4**
Розв'язана задача генерації тексту (для чого застосовувалась нейронна мережа)	Парафразування, підсумовування, спрощення тексту, генерація речень з ключових слів, виправлення граматичних помилок, перенесення стилю	The neural networks are applied to generate text for various applications including paraphrasing, summarization, text simplification, and creating sentences from keywords, all in an unsupervised manner.
Мова генерації тексту	Англійська	The tasks are primarily demonstrated with English examples, though the methodologies are applicable to multiple languages.

Порівняння опису статті [34], виконаному за допомогою іншої ВММ та перевіреному людиною, з результатами в таблиці 2.2 показує, що опис добре узгоджується з результатами роботи як Claude 3 Sonnet, так і GPT-4. Обидві моделі точно визначили тип документу, назву, рік публікації, країни авторів, мету статті, використані архітектури нейронних мереж, метрики якості та розв'язані задачі генерації тексту. Щодо характеристик наборів даних, обидві моделі вказали, що деталі про конкретні набори даних не надаються, за винятком набору даних Quora для парафразування. Вони також зазначили, що стаття зосереджується на зменшенні залежності від розмічених або публічно доступних наборів даних, хоча в різних контекстах обговорюються як розмічені, так і нерозмічені дані.

Таким чином, ВММ можуть точно вилучати ключову інформацію зі статей, хоча іноді може упускати деталі, які явно не вказані в тексті. Для мінімізації ризику неправильного вилучення інформації було виконано перевірку людиною результатів роботи Claude 3 Sonnet. Задля уникнення проблем, пов'язаних із перекладом термінології, відповіді ВММ додатково було затребувано мовою відібраних документів (англійською).

2.4. Оцінка якості

Для оцінки якості процесу відбору та аналізу досліджень у даному огляді були застосовані такі критерії:

1. Чіткість та відповідність критеріїв включення та виключення досліджень меті огляду.

2. Повнота та систематичність пошуку релевантних досліджень у обраних базах даних.

3. Послідовність та відтворюваність процесу відбору досліджень згідно з критеріями включення та виключення.

4. Застосування стандартизованої картки огляду для збору та систематизації даних з відібраних досліджень.

5. Залучення щонайменше двох незалежних дослідників до процесу відбору, аналізу та синтезу даних для мінімізації ризику упередженості.

6. Врахування та опис будь-яких розбіжностей або невизначеностей у процесі відбору та аналізу досліджень.

7. Забезпечення прозорості та відтворюваності процесу огляду шляхом детального опису кожного етапу у звіті.

Дотримання цих критеріїв якості дозволило забезпечити надійність та обґрунтованість результатів і висновків даного систематичного огляду.

PRISMA передбачає наявність у методиці дослідження таких додаткових компонентів:

- *оцінка ризику упередженості у відібраних дослідженнях* не є релевантною через те, що у даному огляді розглядаються різні підходи та методи генерації тексту, а не порівнюються результати окремих досліджень;

- *визначення міри ефекту для кожного результату (або типу результату)* не виконується через те, що цей огляд не має на меті проводити мета-аналіз чи кількісний синтез результатів;

- *опис методів синтезу результатів досліджень*, таких як мета-аналіз, не виконується через те, що огляд не передбачає кількісного синтезу результатів;

- *оцінку ризику упередженості через неповноту подання результатів у публікаціях* не виконується через те, що даний огляд фокусується на описі та класифікації існуючих підходів і методів.

- *оцінки достовірності та надійності результатів*, отриманих із публікацій, не виконується через використання надійних джерел: видань, відібраних Scopus.

Висновки до 2 розділу

У другому розділі описано методику систематичного огляду застосування штучних нейронних мереж для генерації текстового контенту у 2022-2024 рр. Зокрема:

1. Обґрунтовано вибір бази даних Scopus в якості джерела для пошуку публікацій та сформульовано пошуковий запит.

2. Визначено критерії включення та виключення публікацій для систематичного огляду.

3. Описано процес відбору публікацій згідно методики PRISMA, в результаті якого із 248 знайдених документів було відібрано 43 статті для аналізу.

4. Представлено стандартизовану карту огляду публікацій та продемонстровано можливість її автоматизованого заповнення за допомогою великих мовних моделей Claude та GPT-4.

5. Визначено критерії оцінки якості процесу відбору та аналізу досліджень у даному огляді.

Описана методика забезпечує систематичність, відтворюваність та прозорість процесу огляду.

РОЗДІЛ 3
РЕЗУЛЬТАТИ

3.1. Розподіл відібраних документів за роками

У [52] представлені заповнені карти огляду для кожної статті. Результати окремих досліджень не наводяться через те, що цей огляд не має на меті проводити мета-аналіз чи кількісний синтез результатів.

Як видно з рис. 3.1, кількість статей у журналах (ARTICLE) переважає над кількістю матеріалів конференцій (CONFERENCE) протягом 2022-2024 років. У 2022 році кількість документів матераіалів конференції (15) була значно більша за кількість статей у журналах (4), проте у 2023 році спостерігається зростання кількості статей у журналах (16) у порівнянні з матеріалами конференцій (6). За січень та лютий 2024 році наявні лише статті у журналах (2), а матеріали конференцій відсутні. Загалом, за період 2022-2024 років кількість статей у журналах (22) дорівнює кількості матеріалів конференцій (21). Зростання може свідчити більш ґрунтовне висвітлення проблематики у наукових журналах порівняно з матеріалами конференцій за останні роки.

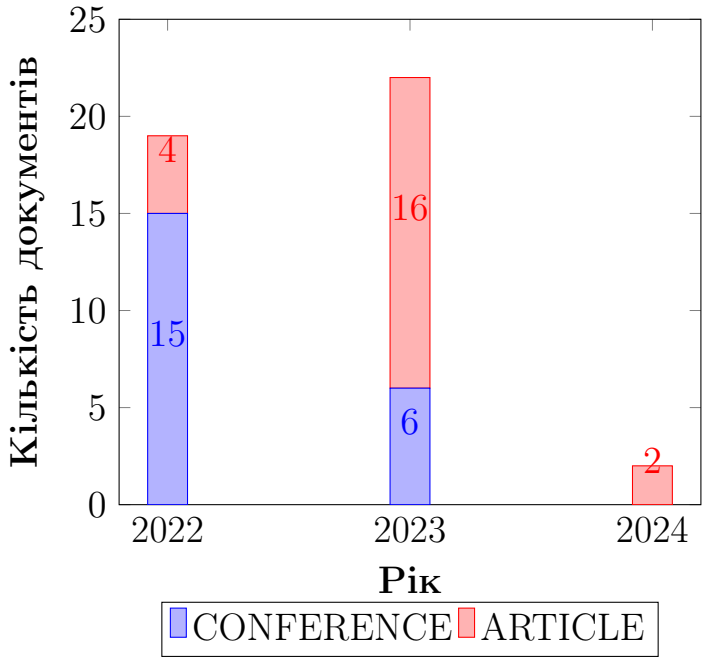

Рис. 3.1. Кількість документів типу CONFERENCE (стаття у матеріалах конференції) та ARTICLE (стаття у журналі) по роках.

3.2. ДП1: Які передові методи глибокого навчання використовуються для генерації тексту в літературі 2022-2024 рр.?

Табл. 3.1 представляє огляд архітектур нейронних мереж, що використовуються для генерації тексту, згідно з даними досліджень 2022-2024 років.

Таблиця 3.1

Архітектури нейронних мереж для генерації тексту.

Архітектура	**Опис**	**Представники**	**Статті**
Традиційні підходи			
RNN (Recurrent Neural Networks)	Рекурентні нейронні мережі, що використовуються для обробки послідовних даних.	–	[1; 4; 8; 21; 25; 34; 43]
LSTM (Long Short-Term Memory)	Варіант RNN, що краще запам'ятовує довгострокові залежності.	–	[1; 6—8; 13; 21; 25; 29; 40; 43; 45]
GRU (Gated Recurrent Unit)	Спрощений варіант LSTM з меншою кількістю параметрів.	–	-
CNN (Convolutional Neural Networks)	Згорткові нейронні мережі, що часто використовуються для обробки зображень.	YOLOv5	[1; 4; 12; 27]
Graph Neural Networks	Моделі, що працюють з графовими структурами даних.	GraphWriter, CGE-LW	[2; 4]
Інноваційні підходи			
Autoencoders	Мережі, які використовують для навчання ефективних кодувань нерозмічених даних	AE, VAE, iVAE, clVAE+ MI, β0.4 VAE, SaVAE, LagVAE	[15; 40; 43]

Продовження на наступній сторінці

Табл. 3.1 – продовження з попередньої сторінки

Архітектура	**Опис**	**Представники**	**Статті**
Transformer	Архітектура, що використовує механізм уваги для обробки послідовних даних.	T5, CodeT5, TrICY, DETR	[2—4; 14; 16; 17; 19; 23; 31; 34; 41; 42; 45; 47; 48; 50; 51]
BERT (Bidirectional Encoder Representations from Transformers)	Модель на основі Transformer, що навчається на великих обсягах нерозміченого тексту.	PubmedBERT, BioLinkBERT, RoBERTa, XLM-RoBERTa	[3; 4; 6; 9; 11; 14; 16; 17; 22; 25; 31; 35; 37; 42; 44; 49]
GPT-2, GPT-3 (Generative Pre-trained Transformer)	Моделі на основі Transformer, що використовуються для генерації тексту.	OPT, Llama, CodeBERT	[1; 3; 6; 8; 9; 14; 16; 18—22; 29; 31; 32; 34; 36; 37; 39—42; 47—49]
Attention-based models	Моделі, що використовують механізм уваги для покращення якості генерованого тексту.	–	[3; 17; 22; 34; 47; 48]
Seq2Seq (Sequence-to-Sequence)	Архітектура, що використовує кодувальник та декодувальник для генерації послідовностей.	S2ST, S2SL, S2SG, S2ST+, D+ Full, DSG	[7; 11; 26; 40; 42; 47; 51]
GAN (Generative Adversarial Networks)	Генеративно-змагальні мережі, що складаються з генератора та дискримінатора.	EGAN, TI-LGAN, DoubAN-Full, WRGAN, CatGAN, SeqGAN, DGSAN	[1; 39; 43]

Продовження на наступній сторінці

Табл. 3.1 – продовження з попередньої сторінки

Архітектура	**Опис**	**Представники**	**Статті**
Memory Networks	Моделі, що використовують зовнішню пам'ять для зберігання та доступу до інформації.	DM-NLG (with memory), MemNNs, Mem2Seq, GLMP	[4; 41]
Diffusion Models	Моделі, що використовують дифузійний процес для генерації тексту.	GENIE, NAT, iNAT, ELMER, MASS, ProphetNet, InsT, CMLM, LevT, BANG, ConstLeven	[45]
Prompt-based models	Моделі, що використовують prompt-engineering донавчання для керування генерацією тексту.	–	[20]

Табл. 3.2 представляє узагальнення підходів до генерації тексту на основі даних табл. 3.1.

Таблиця 3.2

Підходи до генерації тексту.

Категорія	**Статті**
Традиційні підходи	[12; 13; 27]
Інноваційні підходи	[3; 9; 11; 14; 16—20; 22; 23; 26; 31; 32; 35—37; 39; 41; 42; 44; 47—51]
Комбінація традиційних та інноваційних підходів	[1; 2; 4; 6—8; 21; 25; 29; 34; 40; 43; 45]

Серед інноваційних підходів найбільш популярними є використання моделей на основі архітектури Transformer, зокрема GPT-2, GPT-3, BERT та їх варіацій. Ці моделі демонструють високу ефективність у генерації зв'язного та семантично релевантного тексту. Також набувають популярності підходи з використанням механізмів уваги (attention) та контрольо-

ваної генерації тексту (controllable text generation).

Традиційні підходи, хоча і використовуються рідше, все ще знаходять своє застосування у певних задачах, таких як генерація тексту на основі зображень, машинний переклад та інші.

Загалом, спостерігається тенденція до переходу від традиційних підходів до більш інноваційних та ефективних моделей на основі архітектури Transformer та механізмів уваги. Це дозволяє покращити якість генерованого тексту та розширити сферу застосування цих технологій.

Рис. 3.2 показує, що у 2022 та 2023 роках переважають інноваційні підходи до генерації тексту, тоді як традиційні підходи та комбінація підходів зустрічаються рідше. У 2024 році наявні статті, що використовують інноваційний і кобмінований підходи у рівній кількості, проте вибірка за цей рік є неповною, оскільки дані збиралися лише за частину року. Загалом, спостерігається тенденція до збільшення кількості досліджень, що застосовують інноваційні підходи, такі як моделі на основі архітектури Transformer та механізмів уваги.

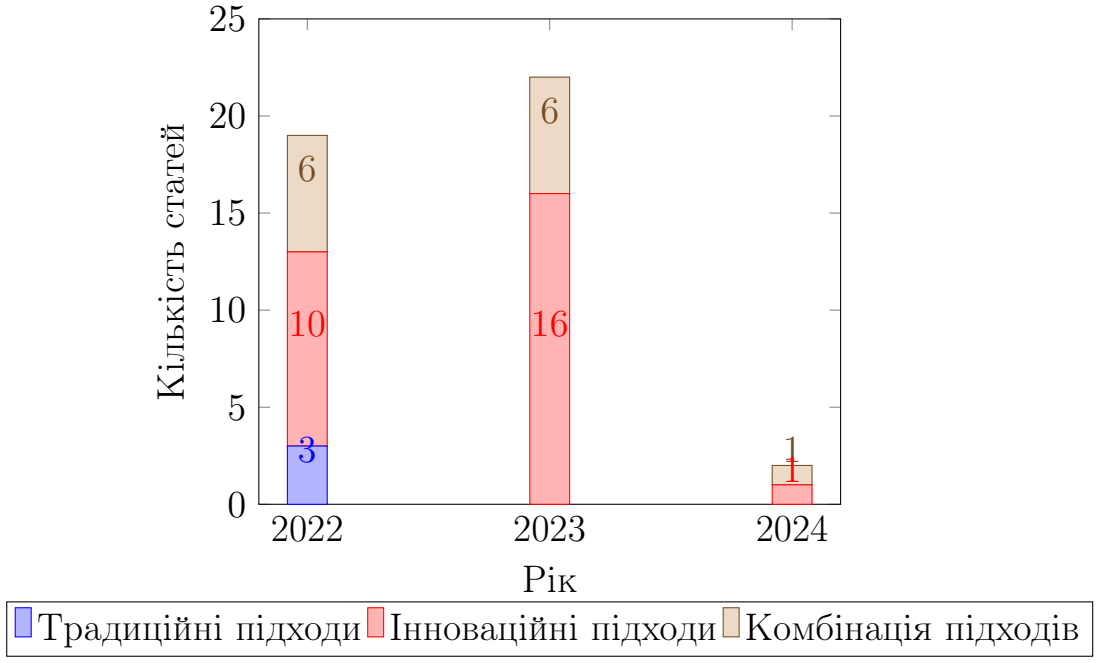

Рис. 3.2. Розподіл статей по роках за категоріями підходів до генерації тексту.

Порівнюючи отримані результати із даними попереднього систематичного огляду [5], можна зробити наступні висновки:

- Традиційні підходи, такі як RNN, LSTM, CNN, все ще використовуються для генерації тексту, але в меншій мірі порівняно з інноваційними підходами.

- Архітектура Transformer та її варіанти (GPT-2, GPT-3, BERT) набули значної популярності у 2022-2024 роках, демонструючи високу ефективність у генерації зв'язного та семантично релевантного тексту.

- З'явилися нові архітектури та підходи, такі як Diffusion Models та Memory Networks models, які не були представлені у попередньому огляді.

- Значна увага приділяється моделям, що використовують механізми уваги (Attention-based models) та контрольованої генерації тексту (Controllable Text Generation).

- Спостерігається тенденція до комбінування традиційних та інноваційних підходів для досягнення кращих результатів у генерації тексту.

- Загалом, у 2022-2024 роках спостерігається перехід від традиційних підходів до більш інноваційних та ефективних моделей на основі архітектури Transformer та механізмів уваги, що дозволяє покращити якість генерованого тексту та розширити сферу застосування цих технологій.

3.3. ДП2: Які нові метрики для оцінювання ефективності моделей генерації тексту в літературі 2022-2024 рр.?

Табл. 3.3 представляє огляд метрик якості, що використовуються для оцінювання генерації тексту. Метрики розділені на дві категорії: human-centred (орієнтовані на людину) та machine-centred (орієнтовані на машину). До human-centred метрик належать Human Evaluation та Turing Test, які передбачають оцінку якості згенерованого тексту людьми-експертами

або тест на здатність моделі генерувати текст, схожий на написаний людиною. Machine-centred метрики включають широкий спектр автоматичних метрик, таких як BLEU, ROUGE, METEOR, Perplexity, Distinct-n, BERTScore та ін. Ці метрики оцінюють різні аспекти якості згенерованого тексту, такі як схожість з еталонним текстом, плавність, змістовність, різноманітність лексики та синтаксису тощо.

Таблиця 3.3

Основні метрики якості для оцінювання генерації тексту.

Метрика якості	**Опис**	**Представники**	**Статті**
Human-centred metrics			
Human Evaluation	Оцінка якості згенерованого тексту людьми-експертами.	–	[4; 8; 21; 25; 29; 31; 36; 37; 39; 51]
Turing Test	Тест на здатність моделі генерувати текст, який неможливо відрізнити від написаного людиною.	–	[29]
Machine-centred metrics			
BLEU	Метрика, що оцінює якість згенерованого тексту шляхом порівняння його з еталонним текстом.	BLEU-1, BLEU-2, BLEU-3, BLEU-4, BLEU-5	[2—4; 6—8; 14; 16; 20; 23; 26; 29; 31; 34; 36; 37; 40; 41; 43; 45; 47—49; 51]
ROUGE	Метрика, що оцінює якість автоматичного реферування тексту.	ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-L	[2—4; 6—8; 11; 14; 16; 20; 23; 26; 29; 35; 36; 41; 45; 47—50]

Продовження на наступній сторінці

Табл. 3.3 – продовження з попередньої сторінки

Метрика якості	**Опис**	**Представники**	**Статті**
METEOR	Метрика, що оцінює якість машинного перекладу.	–	[7; 14; 23; 26; 31; 36; 41; 47; 48; 50]
BERTScore	Метрика, що оцінює якість згенерованого тексту з використанням попередньо навченої моделі BERT.	–	[3; 6; 14; 16; 22; 31; 37; 41]
CIDEr	Метрика, що оцінює якість автоматичного опису зображень, порівнюючи автоматично згенеровані описи з наборами референсних описів.	–	[7; 13; 14; 20; 26; 36; 37; 45]
Perplexity	Метрика, що оцінює якість мовної моделі.	–	[3; 4; 15; 22; 36; 40; 42; 43]
F1-score	Метрика, що оцінює якість класифікації, зокрема в задачах двокласової класифікації.	–	[6; 17; 18; 22; 41; 44]
CHRF++	Метрика, що оцінює якість машинного перекладу, базуючись на збігах символів та n-грам.	–	[2; 31; 37; 50]
Distinct-n	Метрика, що оцінює різноманітність згенерованого тексту.	Dist-1, Dist-2, Dist-3, Dist-4	[3; 4; 40]

Табл. 3.4 надає огляд застосованих у статтях метрик оцінки якості. Більшість досліджень використовують machine-centred метрики для автоматичної оцінки якості згенерованого тексту. Значно менша кількість досліджень застосовує human-centred метрики, що може бути пов'язано з трудомісткістю та суб'єктивністю оцінки якості людьми. Проте, використання human-centred метрик все ще залишається важливим для отримання більш повної та надійної оцінки якості генерації тексту. Деякі дослідження

не застосовують жодних метрик якості, що може бути пов'язано з фокусом на інших аспектах генерації тексту, таких як ефективність чи швидкість роботи моделей.

Таблиця 3.4

Огляд застосованих у статтях метрик оцінки якості.

Метрики якості	**Статті**
Machine-centred	[2; 3; 6; 7; 11—23; 26; 27; 34; 35; 40—45; 47—50]
Human-centred	[21; 25]
Обидві	[4; 8; 29; 31; 36; 37; 39; 51]
Не застосовано	[1; 9; 32]

Використання різноманітних метрик якості є важливим для всебічної оцінки ефективності моделей та підходів до генерації тексту. Комбінування machine-centred та human-centred метрик дозволяє отримати більш надійні та валідні результати оцінювання.

Діаграма на рис. 3.3 показує, що найбільш часто використовуваними метриками якості є BLEU (55.8% статей) та ROUGE (48.8% статей). Також досить поширеною є оцінка якості людьми (Human Evaluation) – вона застосовується у 23.3% статей. Інші метрики, такі як Perplexity, METEOR, BERTScore та Distinct-n, використовуються рідше, але все ще мають значну частку згадувань у статтях. Найменш поширеними є метрики Turing Test, Fluency, Coherence, Diversity, N-gram Overlap та Embedding Similarity, кожна з яких згадується лише в одній статті (2.3%).

Автоматичні метрики якості, такі як BLEU та ROUGE, є найбільш широко використовуваними для оцінки ефективності моделей генерації тексту, тоді як оцінка якості людьми застосовується рідше, але залишається важливим компонентом для отримання більш повної та надійної оцінки якості згенерованого тексту.

Порівнюючи отримані результати з даними попереднього систематичного огляду [5], можна виділити наступні спостереження:

- BLEU та ROUGE залишаються найпопулярнішими метриками для оцінки якості згенерованого тексту як у 2015-2021, так і у 2022-2024 роках.

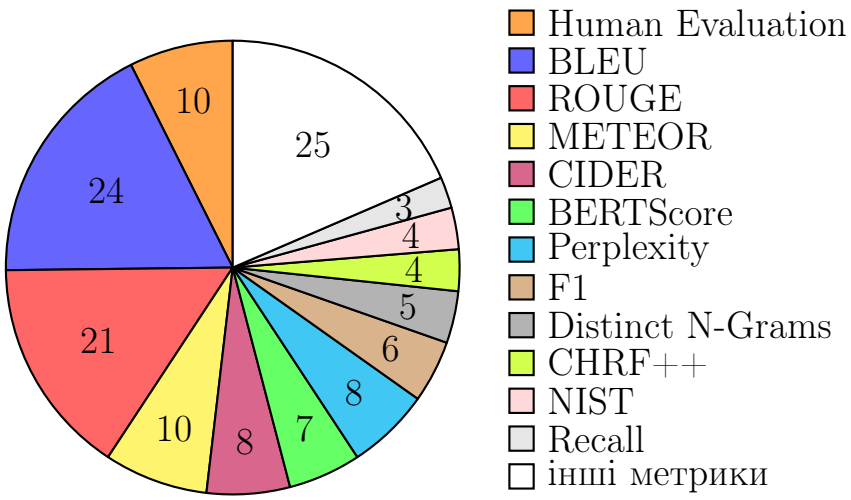

Рис. 3.3. Розподіл метрик якості за кількістю статей, у яких вони згадані.

- Human Evaluation все ще широко застосовується для отримання більш повної та надійної оцінки якості генерації тексту, незважаючи на трудомісткість та суб'єктивність такого підходу.

- У 2022-2024 роках з'явились нові метрики, такі як BERTScore, Fluency, Coherence, Diversity, N-gram Overlap та Embedding Similarity, які не були представлені у попередньому огляді. Це свідчить про активний розвиток методів оцінки якості згенерованого тексту та пошук більш ефективних та інформативних метрик.

- Перплексія (Perplexity) набула більшої популярності у 2022-2024 роках порівняно з попереднім періодом, що може бути пов'язано з її ефективністю в оцінці якості мовних моделей.

- Метрика METEOR, яка оцінює якість машинного перекладу, також частіше використовується у 2022-2024 роках, що може свідчити про зростання інтересу до застосування генерації тексту у задачах машинного перекладу.

- Загалом, спостерігається тенденція до комбінування різних типів метрик (machine-centred та human-centred) для отримання більш надійних та валідних результатів оцінювання ефективності моделей генерації тексту.

Таким чином, порівняння результатів двох оглядів демонструє, що

хоча традиційні метрики, такі як BLEU та ROUGE, залишаються широко використовуваними, у 2022-2024 роках з'являються нові метрики, які враховують різні аспекти якості згенерованого тексту. Це свідчить про активний розвиток методів оцінки якості та пошук більш ефективних та інформативних підходів до оцінювання моделей генерації тексту.

3.4. ДП3: Які набори даних для генерації тексту описані в літературі 2022-2024 рр.?

Табл. 3.5 представляє набори даних, згадані в оглянутих статтях, впорядковані за спаданням кількості згадувань та за алфавітом у випадку однакової кількості згадувань. Набір даних E2E згадується найчастіше – у 7 статтях, за ним слідують XSum (4 статті), CNN/DailyMail (4 статті), CommonGen (4 статті), ToTTo (4 статті), WebNLG (4 статті), WikiBio (3 статті), DDI (2 статті), NIST (2 статті), PUBMed (2 статті), Quora (2 статті), RocStories (2 статті), Snips (2 статті), SST-2 (2 статті), WMT'14 English-German (2 статті), WMT'16 Romanian-English (2 статті) та Yelp (2 статті). Інші набори даних згадуються по 1 разу, впорядковані за появою у огляді.

Таблиця 3.5

Набори даних, згадані в оглянутих статтях.

Назва набору даних	**Статті**
E2E	[16; 20; 25; 36; 41; 48; 51]
CNN/DailyMail (CNN/DM)	[4; 20; 45; 49]
Totto	[14; 26; 47; 51]
CommonGen	[4; 14; 36; 45]
WebNLG	[2; 37; 48; 51]
XSum	[4; 14; 20; 45]
WikiBio	[14; 41; 51]
Abstract Generation Dataset (AGENDA)	[2; 4]
DDI	[4; 9]
NIST	[4; 23]
PubMed	[9; 20]

Продовження на наступній сторінці

Табл. 3.5 – продовження з попередньої сторінки

Назва набору даних	**Статті**
Quora	[4; 34]
ROCStories	[4; 36]
Snips	[16; 42]
SST-2	[18; 49]
WMT'14 English-German	[14; 23]
WMT'16 Romanian-English	[14; 23]
Yelp	[15; 44]
Baidu Tieba	[4]
PersonaChat	[4]
Gigawords	[4]
Yahoo! Answers	[4]
NLPCC	[4]
Tencent	[4]
SQuAD	[4]
ComVE	[4]
$\boldsymbol{\alpha}$NLG-ART	[4]
EntDesc	[4]
VisualStory	[4]
PaperWriting	[4]
Reddit-10M	[4]
EMNLP dialog	[4]
ICLR dialog	[4]
NarrativeQA	[4]
Wizard of Wikipedia (WoW)	[4]
MS-MARCO	[4]
ELI5	[4]
ChangeMyView	[4]
Amazon books	[4]
Foursquare	[4]
Scratch online community comments	[21]
BC5-Chemical	[9]

Продовження на наступній сторінці

Табл. 3.5 – продовження з попередньої сторінки

Назва набору даних	**Статті**
BC5-Disease	[9]
NCBI-Disease	[9]
BC2GM	[9]
JNLPBA	[9]
EBM PICO	[9]
ChemProt	[9]
GAD	[9]
BIOSSES	[9]
HoC	[9]
PubMedQA	[9]
BioASQ	[9]
Logic2Text	[6]
Concadia	[13]
REDIAL	[40]
Custom dataset for Bangla word sign language	[12]
Synthetic dataset	[15]
Penn Treebank	[15]
IWSLT'14 De-En	[14]
WMT16 English-German	[49]
WMT17 English-German	[36]
WMT20	[37]
WMT21	[37]
WMT'14 German-English	[23]
Multi-News	[14]
Java	[14]
Python	[14]
English ATIS	[16]
ViGGO	[16]
TREC	[16]
Korean Weather	[16]
Rest	[16]

Продовження на наступній сторінці

Табл. 3.5 – продовження з попередньої сторінки

Назва набору даних	**Статті**
KLUE-TC	[16]
C4	[17]
M2D2	[17]
Political Slant	[17]
Layoff	[18]
MC	[18]
M&A	[18]
Flood	[18]
Wildfire	[18]
Boston Bombings	[18]
Bohol Earthquake	[18]
West Texas Explosion	[18]
Dublin	[18]
New York City	[18]
WSC	[19]
CBT-CN	[19]
CBT-NE	[19]
Wikihow	[20]
SAMSum	[20]
DART	[20]
Custom dataset composed of tweets labeled with emotions	[39]
AFQMC	[22]
CHIP-STS	[22]
QQP	[22]
MRPC	[22]
ParaNMT-small	[23]
NIST Chinese-English	[23]
GTZAN	[11]
Minions	[43]
Japanimation	[43]

Продовження на наступній сторінці

Табл. 3.5 – продовження з попередньої сторінки

Назва набору даних	**Статті**
WikiArt	[43]
Nottingham	[43]
Lakh MIDI	[43]
TheoryTab	[43]
Poem-5	[43]
Poem-7	[43]
Synthetic date generation dataset	[25]
LDC2020T02 (AMR 3.0 release)	[31]
One Million Urdu News Dataset	[29]
Australian Broadcasting Corporation (ABC) news dataset	[29]
DailyMed drug labels	[35]
COCO Image Captioning	[37]
German and French commercial datasets	[42]
MASSIVE	[42]
Gold-PMB	[7]
Silver-PMB	[7]
numericNLG	[47]
Custom dataset related to text messaging applications	[48]
TweetEval	[49]
AGnews	[49]
QNLI	[49]
IMDB	[49]
CC-News	[49]
WITA	[26]
XWIKIREF	[50]

У проаналізованих дослідженнях використовується широкий спектр наборів даних, що охоплюють різні домени та типи текстів, від відгуків користувачів та новинних статей до медичних і технічних текстів. Це свідчить про активний розвиток та застосування методів генерації текс-

ту у різноманітних сферах.

Табл. 3.6 представляє типи даних, використані в оглянутих статтях, впорядковані за спаданням кількості згадувань. Найчастіше використовуються набори даних, що містять речення – вони згадуються у 26 статтях. У 5 статтях тип даних явно не вказаний. Інші типи даних, такі як абзаци (18 статей), документи (11 статей), питання-відповідь (10 статей), таблиці з описом (9 статей), переклади (7 статей), історії (4 статті), зображення (4 статті) та інші, зустрічаються рідше.

Таблиця 3.6

Типи даних, використані в оглянутих статтях.

Тип даних	**Статті**
Речення	[2; 4; 9; 13—17; 19—22; 25; 26; 29; 31; 36; 39; 40; 42—45; 49—51]
Абзац	[2; 4; 9; 13—17; 20; 26; 37; 40; 42—45; 49; 50]
Документ	[4; 7; 9; 14; 16; 17; 35; 43—45; 49]
Питання-відповідь	[2; 4; 9; 14—16; 19; 22; 34; 49]
Таблиці з описом	[6; 18; 25; 29; 36; 41; 47; 48; 51]
Переклади	[14; 23; 29; 36; 37; 49; 51]
Історії	[4; 29; 36; 51]
Зображення	[13; 27; 37; 43]
Аудіофайли	[11; 43]
Відеокліпи	[12]
Комп'ютерні програми	[14]
Не вказаний	[1; 3; 8; 32]

Переважання наборів даних з реченнями може бути пов'язане з тим, що багато завдань генерації тексту, таких як машинний переклад, парафразування, відповіді на запитання тощо, часто працюють на рівні речень. Водночас, наявність різноманітних типів даних, включаючи абзаци, документи, зображення, музику та інші, свідчить про те, що методи генерації тексту можуть застосовуватись до широкого спектру задач та доменів.

Табл. 3.7 представляє типи розмітки даних, використані в оглянутих статтях, впорядковані за спаданням кількості згадувань. Найчастіше використовуються розмічені набори даних – вони згадуються у 22 статтях.

У 20 статтях тип розмітки явно не вказаний. У 5 статтях використовуються нерозмічені дані. У 4 статтях використовуються як розмічені, так і нерозмічені дані.

Таблиця 3.7

Типи розмітки даних, використані в оглянутих статтях.

Тип розмітки	**Статті**
Розмічені дані	[6; 7; 9; 11—15; 18; 23; 26; 29; 34; 35; 39; 41; 42; 44; 47; 48; 50; 51]
Нерозмічені дані	[9; 21; 34; 42; 44]
Не вказано	[1—4; 8; 16; 17; 19; 20; 22; 25; 27; 31; 32; 36; 37; 40; 43; 45; 49]

Переважання розмічених наборів даних може бути пов'язане з тим, що багато завдань генерації тексту, особливо ті, що використовують контрольовані підходи або вимагають відповідності певним шаблонам чи структурам, потребують розмічених даних для навчання моделей. Розмітка може включати такі елементи, як частини мови, синтаксичні структури, семантичні ролі, теги для контрольованої генерації тощо.

Водночас, наявність досліджень, що використовують нерозмічені дані або комбінацію розмічених та нерозмічених даних, свідчить про активний розвиток методів навчання без учителя та напівавтоматичного навчання в галузі генерації тексту. Ці підходи дозволяють використовувати великі обсяги нерозмічених текстових даних для попереднього навчання моделей та покращення їх здатності до генерації зв'язного та змістовного тексту.

Табл. 3.8 представляє якість даних, використаних в оглянутих статтях, впорядковані за спаданням кількості згадувань. У 28 статтях якість даних явно не вказана. У 12 статтях використовуються попередньо опрацьовані дані, тоді як у 10 статтях – сирі дані. У 4 статтях використовуються як попередньо опрацьовані, так і сирі дані.

Попередньо опрацьовані дані зазвичай проходять етапи очищення, нормалізації, токенізації, а іноді й додаткової розмітки перед використанням у навчанні моделей. Це дозволяє покращити якість та консистентність даних, а також полегшити процес навчання. Прикладами попередньо опрацьованих даних можуть бути набори даних, отримані з існуючих корпусів

або баз даних, які вже пройшли певну обробку.

Таблиця 3.8

Якість даних, використаних в оглянутих статтях.

Якість даних	**Статті**
Попередньо опрацьовані	[6; 7; 12—15; 34; 41; 42; 48; 50; 51]
Сирі	[2; 7; 11; 21; 29; 35; 37; 41; 42; 51]
Не вказано	[1; 3; 4; 8; 9; 16—20; 22; 23; 25—27; 31; 32; 36; 39; 40; 43—45; 47; 49]

Сирі дані, з іншого боку, – це дані, отримані безпосередньо з реальних джерел, таких як веб-сторінки, соціальні мережі, необроблені тексти тощо. Вони можуть містити шум, некоректне форматування, помилки та інші артефакти. Використання сирих даних може бути корисним для навчання моделей, які мають бути стійкими до реальних умов та здатними обробляти неструктуровані дані.

Відсутність інформації про якість даних у значній частині проаналізованих статей може свідчити про те, що автори не приділяють достатньої уваги цьому аспекту або вважають його менш важливим для дослідження. Водночас, якість даних є критичним фактором, що впливає на ефективність та узагальнюваність моделей генерації тексту, тому варто приділяти більше уваги опису та аналізу якості використаних даних у майбутніх дослідженнях.

Порівнюючи результати огляду 2022-2024 років із попереднім оглядом [5], можна зробити наступні висновки:

- У 2022-2024 роках з'явилися нові набори даних, такі як XWIKIREF, DailyMed, numericNLG, WITA, DIST-ToTTo, які не були представлені в попередньому огляді. Це свідчить про активний розвиток ресурсів для дослідження та застосування методів генерації тексту.

- Набори даних E2E, WikiBio, ToTTo, CommonGen, CNN/DailyMail та XSum залишаються популярними і широко використовуються в дослідженнях як у 2015-2021, так і в 2022-2024 роках.

- Спостерігається тенденція до використання більш різноманітних типів даних, таких як таблиці з описом, зображення, музика, переклади, питання-відповідь, відеокліпи та комп'ютерні програми, на додаток до традиційних типів, як-от речення, абзаци та документи.

- Розмічені дані залишаються найбільш широко використовуваними, але спостерігається зростання інтересу до використання нерозмічених даних та комбінації розмічених і нерозмічених даних для навчання моделей генерації тексту.

- Хоча якість даних є критичним фактором, що впливає на ефективність моделей, у значній частині досліджень 2022-2024 років цей аспект не висвітлюється, що може свідчити про необхідність приділяти більше уваги опису та аналізу якості використаних даних у майбутніх дослідженнях.

Таким чином, порівняння результатів двох оглядів демонструє, що набори даних для генерації тексту продовжують активно розвиватися, охоплюючи нові домени та типи даних. Водночас, деякі популярні набори даних залишаються актуальними та широко використовуваними в дослідженнях. Спостерігається тенденція до використання більш різноманітних типів даних та зростання інтересу до нерозмічених даних і комбінованих підходів. Проте, опис якості даних все ще потребує більшої уваги в майбутніх дослідженнях для забезпечення надійності та відтворюваності результатів.

3.5. ДП4: Які нові застосування генерації тексту описані в літературі 2022-2024 рр.?

Табл. 3.9 відображає застосування генерації тексту, знайдені у проаналізованих статтях, впорядковані за спаданням кількості посилань. Найбільш поширеними застосуваннями є генерація анотацій (8 статей), машинний переклад (8 статей), генерація тексту з таблиць (5 статей), перефразування та доповнення даних (по 4 статті). Інші застосування, такі як контрольована генерація тексту, генерація тексту на основі зображень, генерація тексту з графів зань тощо згадуються у меншій кількості статей.

Таблиця 3.9

Застосування генерації тексту.

Застосування	**Статті**
Генерація тексту з таблиць (Table-to-Text Generation)	[6; 25; 36; 41; 47]
Генерація тексту з графів знань (Text Generation from Knowledge Graphs)	[2; 4]
Контрольована генерація тексту (Controllable Text Generation)	[3; 16; 25]
Генерація медичних текстів (Medical Text Generation)	[9; 35]
Парафразування (Paraphrasing)	[4; 23; 34; 42]
Генерація тексту на основі зображень (Image-based Text Generation)	[13; 37; 43]
Генерація анотацій (Text Summarization)	[4; 14; 20; 34; 37; 45; 49; 50]
Генерація емоційно забарвленого тексту (Emotional Text Generation)	[21; 39]
Генерація відповідей на запитання (Question Answering)	[4; 40]
Генерація музичних текстів (Music Text Generation)	[11; 43]
Машинний переклад (Machine Translation)	[4; 12; 14; 15; 23; 34; 36; 37]
Доповнення даних (Data Augmentation)	[3; 7; 18; 44]
Генерація сценаріїв (Script Generation)	[4; 43]
Генерація новинних заголовків (News Headline Generation)	[29]
Генерація технічної документації (Technical Documentation Generation)	[8]
Кібербезпека (Cybersecurity)	[49]
Генерація енциклопедичних статей (Encyclopedic Text Generation)	[50]
Генерація тексту з структурованих даних (Data-to-Text Generation)	[3; 26; 41; 48; 51]
Переклад жестової мови в текст (Sign Language to Text Translation)	[12]

Рисунок 3.4 візуалізує наведені у таблиці 3.9 застосування генерації тексту у вигляді діаграми. На діаграмі чітко видно переважання застосувань генерації тексту з таблиць, графів знань, контрольованої генерації тексту та генерації медичних текстів у порівнянні з іншими напрямками.

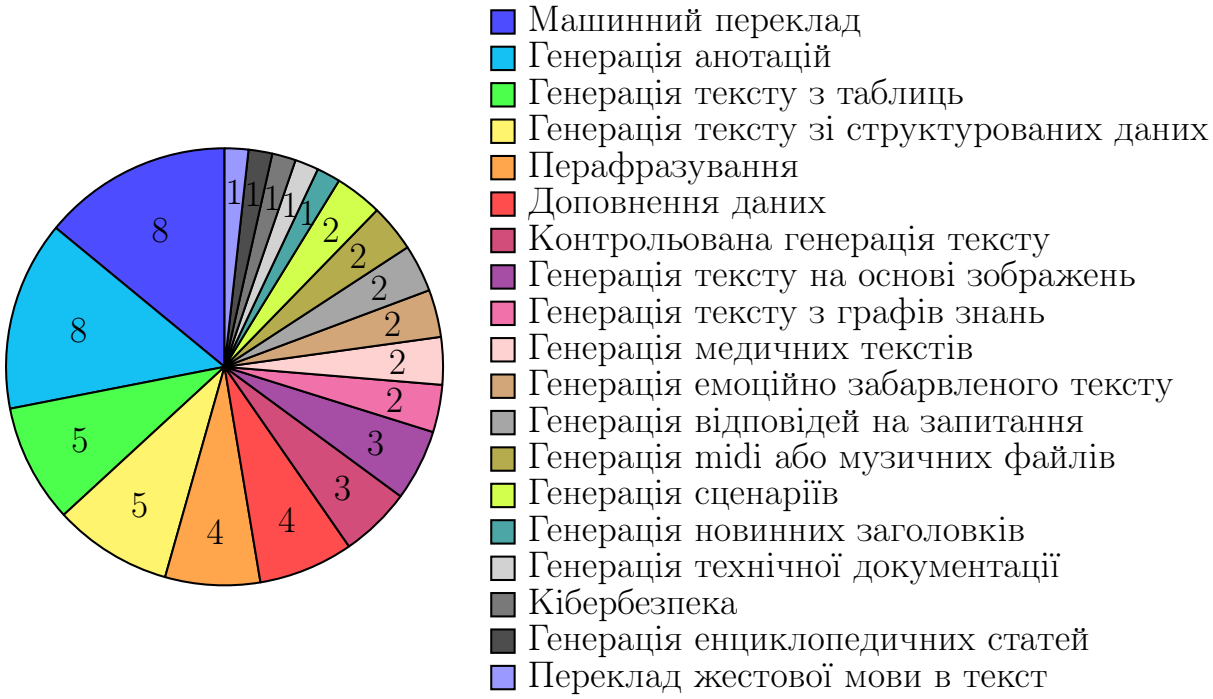

Рис. 3.4. Застосування генерації тексту у проаналізованих статтях 2022-2024 рр.

Аналіз застосувань генерації тексту демонструє широкий спектр можливостей використання цієї технології у різних галузях, від обробки структурованих даних до створення емоційно забарвлених текстів та переклад жестової мови в текст. Розвиток нових методів та архітектур нейронних мереж відкриває нові перспективи для подальшого розширення сфер застосування генерації тексту.

Порівнюючи результати огляду 2022-2024 років із попереднім оглядом [5], можна виділити наступні спостереження:

- Машинний переклад (Machine Translation) та генерація анотацій (Text Summarization) набули більшої популярності у 2022-2024 роках порівняно з попереднім періодом. Проте у 2022-2024 роках до них додалися генерація текстів з таблиць та структурованих даних, що може вказувати на зростання інтересу до обробки структурованої інформації методами генерації тексту.

- Контрольована генерація тексту (Controllable Text Generation) також стала більш поширеною, що свідчить про зростання інтересу до методів, які дозволяють керувати процесом генерації тексту та отримувати більш релевантні та якісні результати.

- Генерація медичних текстів (Medical Text Generation) з'явилася як новий напрямок застосування генерації тексту у 2022-2024 роках, що може бути пов'язано з активним розвитком методів обробки медичних даних та потребою в автоматизації створення медичної документації.

- З'явилися нові застосування, такі як генерація емоційно забарвлених текстів (Emotional Text Generation), генерація енциклопедичних статей (Encyclopedic Text Generation), генерація технічної документації (Technical Documentation Generation) та переклад жестової мови в текст (Sign Language to Text Translation), що свідчить про розширення сфер використання генерації тексту.

- Парафразування та доповнення даних залишаються актуальними застосуваннями генерації тексту як у 2015-2021, так і в 2022-2024 роках.

- Деякі застосування, які були популярними у попередньому огляді, такі як генерація поезії, діалогових систем, класифікація текстів, topic modeling, у новому огляді не фігурують серед найбільш згадуваних. Це може бути пов'язано зі зміною фокусу досліджень та появою нових перспективних напрямків.

- Загалом, спостерігається тенденція до зростання різноманітності застосувань генерації тексту порівняно з попереднім, що свідчить про активний розвиток цього напрямку досліджень та розширення можливостей використання генеративних моделей для вирішення прикладних завдань у різних предметних областях.

Таким чином, порівняння результатів двох оглядів демонструє, що сфера застосування генерації тексту продовжує активно розширюватися, охоплюючи нові галузі та напрямки. Популярність таких застосувань, як генерація тексту з таблиць та графів знань, контрольована генерація тексту

та генерація медичних текстів, свідчить про зростання інтересу до методів, які дозволяють ефективно обробляти структуровані дані та отримувати більш релевантні та якісні результати. Водночас, традиційні застосування, такі як парафразування, генерація анотацій та машинний переклад, залишаються актуальними та широко використовуваними в дослідженнях.

3.6. ДП5: Які природні мови використовуються для генерації тексту в літературі 2022-2024 рр.?

Табл. 3.10 представляє розширене річне зведення мов генерації тексту за 2022-2024 роки. Англійська мова є найбільш широко використовуваною, з 38 статтями, що охоплюють усі три роки. Для генерації англійських текстів використовуються різноманітні архітектури нейронних мереж, включаючи Transformer, BERT, GPT-2, GPT-3, RNN, LSTM, CNN, GAN та Seq2Seq.

Німецька мова представлена у 5 статтях з використанням архітектури GAN (Conditional GAN, StyleGAN, DCGAN). Китайська мова представлена у 4 статтях з використанням Graph Neural Networks та архітектури B2T. Бенгальська мова представлена у 2 статтях (по одній у 2022 та 2023 роках), присвячених розпізнаванню з використанням CNN та YOLO. Румунська мова представлена у 2 статтях (по одній у 2022 та 2023 роках) з використанням архітектур DCGAN та BART. Французька, урду, шекспірівська англійська та корейська мови згадуються по одній статті кожна, з використанням різних архітектур, таких як Conditional GAN, StyleGAN, DCGAN та GPT-2.

У 2023 році з'являється дослідження [50], присвячене генерації текстів одразу кількома індійськими мовами (хінді, малаялам, маратхі, орія, панджабі та тамільська) з використанням архітектур HipoRank, mBART та mT5.

Таблиця 3.10

Розширене річне зведення мов генерації тексту.

Мова	**2022**	**2023**	**2024**	**Разом**	**Архітектури**
Англійська	17 [1; 7; 11—15; 19; 20; 25; 34; 36; 37; 39; 42; 47; 51]	19 [2; 3; 9; 16—18; 21—23; 26; 29; 31; 35; 40; 41; 44; 45; 49; 50]	2 [6; 48]	38	Transformer, BERT, GPT-2, GPT-3, RNN, LSTM, CNN, GAN, Seq2Seq
Німецька	3 [14; 37; 42]	2 [23; 49]	–	5	Conditional GAN, StyleGAN, DCGAN
Китайська	1 [37]	3 [8; 23; 43]	–	4	Graph Neural Networks, B2T
Французька	1 [42]	–	–	1	Conditional GAN, StyleGAN, DCGAN
Бенгальська	1 [12]	1 [50]	–	2	CNN, YOLO, mBART
Урду	–	1 [29]	–	1	GPT-2
Хінді, малаялам, маратхі, орія, панджабі, тамільська	–	1 [50]	–	1	HipoRank, mBART, mT5
Шекспірівська англійська	1 [43]	–	–	1	Modified DCGAN
Румунська	1 [14]	1 [23]	–	2	DCGAN, BART

Продовження на наступній сторінці

Табл. 3.10 – продовження з попередньої сторінки

Мова	**2022**	**2023**	**2024**	**Разом**	**Архітектури**
Корейська	–	1 [16]	–	1	Modified DCGAN

Порівнюючи результати огляду 2022-2024 років із попереднім оглядом [5], можна виділити наступні спостереження:

- Англійська мова залишається найбільш широко використовуваною для генерації тексту як у 2015-2021, так і в 2022-2024 роках. Проте, спостерігається тенденція до збільшення кількості досліджень, присвячених іншим мовам, особливо мовам з обмеженими ресурсами.

- У 2022-2024 роках з'явилися дослідження, присвячені генерації текстів мовами, які не були представлені в попередньому огляді, такими як урду, хінді, малаялам, маратхі, орія, панджабі та тамільська. Це свідчить про зростаючий інтерес до розробки моделей генерації тексту для різноманітних мов.

- Дослідження [50] демонструє можливість генерації текстів одразу кількома індійськими мовами з використанням сучасних архітектур, таких як HipoRank, mBART та mT5, що не було представлено в попередньому огляді.

- Для генерації текстів різними мовами використовуються як традиційні архітектури (RNN, LSTM, CNN), так і більш сучасні підходи, такі як Transformer, BERT, GPT-2, GPT-3, GAN та Graph Neural Networks.

- Загалом, спостерігається тенденція до розширення спектру мов, для яких розробляються моделі генерації тексту, та використання більш різноманітних архітектур нейронних мереж для цієї задачі.

Таким чином, порівняння результатів двох оглядів демонструє, що хоча англійська мова залишається домінуючою в дослідженнях з генерації тексту, спостерігається зростаючий інтерес до розробки моделей для інших

мов, особливо мов з обмеженими ресурсами. Поява досліджень, присвячених генерації текстів такими мовами, як урду, хінді, малаялам, маратхі, орія, панджабі та тамільська, свідчить про розширення можливостей застосування генерації тексту для різноманітних мов. Крім того, використання сучасних архітектур нейронних мереж, таких як Transformer, BERT, GPT-2, GPT-3, GAN та Graph Neural Networks, дозволяє покращити якість та ефективність генерації тексту для різних мов.

Порівнюючи розподіл мов у старому та новому оглядах з розподілом мов за кількістю моделей на Hugging Face [28], можна зробити наступні спостереження:

- Англійська мова домінує в усіх трьох розподілах. У старому та новому оглядах вона є найбільш широко використовуваною для генерації тексту, а на Hugging Face для неї доступно найбільше моделей (51738). Це свідчить про значну увагу дослідників та розробників до англійської мови та наявність великої кількості ресурсів для неї.

- Китайська мова посідає друге місце за кількістю моделей на Hugging Face (4546) та згадується в кількох статтях у новому огляді. Це вказує на зростаючий інтерес до генерації тексту китайською мовою та розвиток відповідних ресурсів.

- Такі мови, як французька, іспанська, російська та німецька, мають значну кількість моделей на Hugging Face (від 2326 до 4049), але рідше згадуються в оглядах. Це може свідчити про те, що, незважаючи на наявність ресурсів для цих мов, дослідження генерації тексту для них не так широко представлені в літературі.

- Мови з обмеженими ресурсами, такі як бенгальська, урду, арабська та хінді, згадуються в новому огляді, що свідчить про зростаючий інтерес до розробки моделей генерації тексту для цих мов. Однак кількість доступних моделей на Hugging Face для цих мов значно менша порівняно з англійською (від 670 до 1674).

- На Hugging Face представлено значно більше мов (більше 200), ніж згадується в оглядах. Це вказує на те, що дослідження генерації текс-

ту охоплюють лише частину мов, для яких доступні моделі та ресурси.

- Деякі мови, такі як японська, корейська, індонезійська та арабська, мають значну кількість моделей на Hugging Face (від 1674 до 2920), але рідко згадуються в оглядах. Це може свідчити про потенціал для подальших досліджень генерації тексту цими мовами.

Порівняння розподілів мов показує, що, незважаючи на домінування англійської мови в дослідженнях та наявних ресурсах, спостерігається зростаючий інтерес до генерації тексту іншими мовами, особливо з обмеженими ресурсами. Однак кількість доступних моделей та ресурсів для цих мов все ще значно менша порівняно з англійською. Крім того, наявність великої кількості моделей для деяких мов на Hugging Face, які рідко згадуються в оглядах, вказує на потенціал для подальших досліджень та розробок у цій галузі.

Висновки до 3 розділу

У третьому розділі представлено результати систематичного огляду застосування штучних нейронних мереж для генерації текстового контенту у 2022-2024 рр. та проведено порівняння з результатами попереднього огляду за 2015-2021 роки. Основні висновки можна узагальнити наступним чином:

1. Спостерігається тенденція до збільшення кількості статей у наукових журналах порівняно з матеріалами конференцій, що може свідчити про більш ґрунтовне висвітлення проблематики генерації тексту в журналах.

2. Серед передових методів глибокого навчання для генерації тексту найбільш популярними є моделі на основі архітектури Transformer, такі як GPT-2, GPT-3, BERT та їх варіації. Також набувають популярності підходи з використанням механізмів уваги та контрольованої генерації тексту. Загалом, спостерігається перехід від традиційних підходів до більш інноваційних та ефективних моделей.

3. Серед метрик для оцінювання ефективності моделей генерації тексту найбільш широко використовуються BLEU та ROUGE, а також оцінка якості людьми (Human Evaluation). У 2022-2024 роках з'явилися нові метрики, такі як BERTScore, Fluency, Coherence, Diversity, N-gram Overlap та Embedding Similarity, що свідчить про активний розвиток методів оцінки якості згенерованого тексту.

4. Набори даних для генерації тексту продовжують активно розвиватися, охоплюючи нові домени та типи даних. Спостерігається тенденція до використання більш різноманітних типів даних (таблиці з описом, зображення, музика, переклади тощо) та зростання інтересу до нерозмічених даних і комбінованих підходів.

5. Сфера застосування генерації тексту продовжує активно розширюватися, охоплюючи нові галузі та напрямки. Популярність таких застосувань, як генерація тексту з таблиць та графів знань, контрольована генерація тексту та генерація медичних текстів, свідчить про зростання інтересу до методів, які дозволяють ефективно обробляти структуровані дані та отримувати більш релевантні та якісні результати.

6. Хоча англійська мова залишається домінуючою в дослідженнях з генерації тексту, спостерігається зростаючий інтерес до розробки моделей для інших мов, особливо мов з обмеженими ресурсами. Використання сучасних архітектур нейронних мереж дозволяє покращити якість та ефективність генерації тексту для різних мов.

ВИСНОВКИ

У даній роботі було проведено систематичний огляд застосування штучних нейронних мереж для генерації текстового контенту у 2022-2024 роках. Основною метою дослідження було виявити та узагальнити сучасні тенденції, підходи та методи в цій області, а також порівняти отримані результати з даними попереднього систематичного огляду за 2015-2021 роки. На основі проведеного аналізу можна зробити такі висновки:

1. Серед передових методів глибокого навчання для генерації тексту домінують моделі на основі архітектури Transformer, такі як GPT-2, GPT-3, BERT та їх модифікації. Також набувають популярності підходи з використанням механізмів уваги та контрольованої генерації тексту. Загалом спостерігається тенденція до переходу від традиційних методів до більш інноваційних та ефективних моделей.

2. Найбільш широко використовуваними метриками для оцінювання ефективності моделей генерації тексту залишаються BLEU та ROUGE, а також оцінка якості людьми (Human Evaluation). Водночас у 2022-2024 роках з'явилися нові метрики, такі як BERTScore, Fluency, Coherence, Diversity, N-gram Overlap та Embedding Similarity, що свідчить про активний розвиток методів оцінки якості згенерованого тексту.

3. Набори даних для генерації тексту продовжують активно розвиватися, охоплюючи нові домени та типи даних. Спостерігається тенденція до використання більш різноманітних типів даних (таблиці з описом, зображення, музика, переклади тощо) та зростання інтересу до нерозмічених даних і комбінованих підходів.

4. Сфера застосування генерації тексту розширюється, охоплюючи нові галузі та напрямки. Популярними стають такі застосування, як генерація тексту з таблиць та графів знань, контрольована генерація тексту та генерація медичних текстів, що свідчить про зростання інтересу до методів, які дозволяють ефективно обробляти структуровані дані та отримувати більш релевантні та якісні результати.

5. Хоча англійська мова залишається домінуючою в дослідженнях з генерації тексту, спостерігається зростаючий інтерес до розробки моделей для інших мов, особливо мов з обмеженими ресурсами. Використання сучасних архітектур нейронних мереж дозволяє покращити якість та ефективність генерації тексту для різних мов.

Отримані результати демонструють активний розвиток та еволюцію методів генерації текстового контенту з використанням штучних нейронних мереж у 2022-2024 роках порівняно з попереднім періодом. Застосування нових архітектур, підходів та метрик дозволяє покращити якість, релевантність та ефективність генерованого тексту, а також розширити сферу застосування цих технологій.

Результати даного систематичного огляду можуть бути корисними для дослідників та практиків у галузі обробки природної мови та штучного інтелекту, оскільки вони надають актуальну інформацію про сучасні тенденції та напрямки розвитку генерації текстового контенту. Отримані висновки можуть бути використані для вибору відповідних методів, архітектур, метрик та наборів даних при розробці нових моделей та систем генерації тексту, а також для визначення перспективних напрямків подальших досліджень у цій області.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. A Brief History of Deep Learning-Based Text Generation / A. Bas [та ін.] // Proceedings of the International Conference on Computer and Applications, ICCA 2022 - Proceedings / за ред. J. M. Alja'Am [та ін.]. — Institute of Electrical, Electronics Engineers Inc., 2022. — DOI: 10.1109/ICCA56443.2022.10039545.

2. A quantum-like approach for text generation from knowledge graphs / J. Zhu [та ін.] // CAAI Transactions on Intelligence Technology. — 2023. — DOI: 10.1049/cit2.12178.

3. A Survey of Controllable Text Generation Using Transformer-based Pre-trained Language Models / H. Zhang [та ін.]. — 2023. — DOI: 10.1145/3617680.

4. A Survey of Knowledge-enhanced Text Generation / W. Yu [та ін.] // ACM Computing Surveys. — 2022. — Т. 54, 11 s. — DOI: 10.1145/3512467.

5. A Systematic Literature Review on Text Generation Using Deep Neural Network Models / N. Fatima [та ін.] // IEEE Access. — 2022. — Т. 10. — С. 53490—53503. — DOI: 10.1109/ACCESS.2022.3174108.

6. *Alonso I.*, *Agirre E.* Automatic Logical Forms improve fidelity in Table-to-Text generation // Expert Systems with Applications. — 2024. — Т. 238. — DOI: 10.1016/j.eswa.2023.121869.

7. *Amin M. S.*, *Mazzei A.*, *Anselma L.* Towards Data Augmentation for DRS-to-Text Generation // CEUR Workshop Proceedings / за ред. D. Nozza, L. Passaro, M. Polignano. — 2022. — Т. 3287. — С. 141—152.

8. An automatic text generation algorithm of technical disclosure for catenary construction based on knowledge element model / J. Wu [та ін.] // Advanced Engineering Informatics. — 2023. — Т. 56. — С. 101913. — DOI: https://doi.org/10.1016/j.aei.2023.101913. — URL: https://www.sciencedirect.com/science/article/pii/S1474034623000411.

9. An extensive benchmark study on biomedical text generation and mining with ChatGPT / Q. Chen [та ін.] // Bioinformatics. — 2023. — Т. 39, № 9. — DOI: 10.1093/bioinformatics/btad557.

10. *awesomegpts.ai*. Scholar GPT. — 2024. — URL: https://chatgpt.com/g/g-kZ0eYXlJe-scholar-gpt?oai-dm=1.

11. *Chu X.* Feature Extraction and Intelligent Text Generation of Digital Music // Computational Intelligence and Neuroscience. — 2022. — Т. 2022. — DOI: 10.1155/2022/7952259.

12. Computer Vision-Based Bengali Sign Language To Text Generation / T. Tazalli [та ін.] // 5th IEEE International Image Processing, Applications and Systems Conference, IPAS 2022. — Institute of Electrical, Electronics Engineers Inc., 2022. — DOI: 10.1109/IPAS55744.2022.10052928.

13. Concadia: Towards Image-Based Text Generation with a Purpose / E. Kreiss [та ін.] // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022 / за ред. Y. Goldberg, Z. Kozareva, Y. Zhang. — Association for Computational Linguistics (ACL), 2022. — С. 4667—4684.

14. CONT: Contrastive Neural Text Generation / C. An [та ін.] // Advances in Neural Information Processing Systems. Т. 35 / за ред. S. Koyejo [та ін.]. — Neural information processing systems foundation, 2022.

15. Contrastive Latent Variable Models for Neural Text Generation / Z. Teng [та ін.] // Proceedings of Machine Learning Research. Т. 180 / за ред. J. Cussens, K. Zhang. — ML Research Press, 2022. — С. 1928—1938.

16. Controllable Text Generation Using Semantic Control Grammar / H. Seo [та ін.] // IEEE Access. — 2023. — Т. 11. — С. 26329—26343. — DOI: 10.1109/ACCESS.2023.3252017.

17. Controlled Text Generation with Natural Language Instructions / W. Zhou [та ін.] // Proceedings of Machine Learning Research. Т. 202 / за ред. A. Krause [та ін.]. — ML Research Press, 2023. — С. 42602—42613.

18. Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers / M. Bayer [та ін.] // International Journal of Machine Learning and Cybernetics. — 2023. — Т. 14, № 1. — С. 135—150. — DOI: 10.1007/s13042-022-01553-3.

19. DFX: A Low-latency Multi-FPGA Appliance for Accelerating Transformer-based Text Generation / S. Hong [та ін.] // Proceedings of the Annual International Symposium on Microarchitecture, MICRO. 2022—October. — IEEE Computer Society, 2022. — С. 616—630. — DOI: 10.1109/MICRO56248.2022.00051.

20. Discourse-Aware Soft Prompting for Text Generation / M. Ghazvininejad [та ін.] // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022 / за ред. Y. Goldberg, Z. Kozareva, Y. Zhang. — Association for Computational Linguistics (ACL), 2022. — С. 4570—4589.

21. *Du H.*, *Xing W.*, *Pei B.* Automatic text generation using deep learning: providing large-scale support for online learning communities // Interactive Learning Environments. — 2023. — Т. 31, № 8. — С. 5021—5036. — DOI: 10.1080/10494820.2021.1993932.

22. Enhancing Text Generation with Cooperative Training / T. Wu [та ін.] // Frontiers in Artificial Intelligence and Applications / за ред. K. Gal [та ін.]. — 2023. — Т. 372. — С. 2704—2711. — DOI: 10.3233/FAIA230579.

23. Explicit Syntactic Guidance for Neural Text Generation / Y. Li [та ін.] // Proceedings of the Annual Meeting of the Association for Computational Linguistics. Т. 1. — Association for Computational Linguistics (ACL), 2023. — С. 14095—14112.

24. *Ganegedara T.* Natural Language Processing with TensorFlow: Teach language to machines using Python's deep learning library. — Birmingham – Mumbai : Packt Publishing, 2018. — URL: https://cdnpdf.com/pdf-38273-natural-language-processing-with-tensorflow-teach-language-to-machines-using-python%27s-deep-learning.

25. GenNI: Human-AI Collaboration for Data-Backed Text Generation / H. Strobelt [та ін.] // IEEE Transactions on Visualization and Computer

Graphics. — 2022. — Т. 28, № 1. — С. 1106—1116. — DOI: 10.1109/TVCG.2021.3114845.

26. *Gong H.*, *Feng X.*, *Qin B.* Quality Control for Distantly-Supervised Data-to-Text Generation via Meta Learning // Applied Sciences (Switzerland). — 2023. — Т. 13, № 9. — DOI: 10.3390/app13095573.

27. *Hanafi A.*, *Bouhorma M.*, *Elaachak L.* Machine Learning-Based Augmented Reality For Improved Text Generation Through Recurrent Neural Networks // Journal of Theoretical and Applied Information Technology. — 2022. — Т. 100, № 2. — С. 518—530.

28. *Hugging Face*. Languages. — 06.2024. — URL: https://huggingface.co/languages.

29. Improving news headline text generation quality through frequent POS-Tag patterns analysis / N. Fatima [та ін.] // Engineering Applications of Artificial Intelligence. — 2023. — Т. 125. — DOI: 10.1016/j.engappai.2023.106718.

30. Introducing the next generation of Claude. — 2024. — URL: https://www.anthropic.com/news/claude-3-family.

31. Investigating the Effect of Relative Positional Embeddings on AMR-to-Text Generation with Structural Adapters / S. Montella [та ін.] // EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference. — Association for Computational Linguistics (ACL), 2023. — С. 727—736.

32. *Koplin J. J.* Dual-use implications of AI text generation // Ethics and Information Technology. — 2023. — Т. 25, № 2. — DOI: 10.1007/s10676-023-09703-z.

33. large language models - Google Trends. — 12.2023. — URL: https://trends.google.com/trends/explore?date=2022-01-01%202023-12-21&q=large%20language%20models&hl=en.

34. *Mou L.* Search and learning for unsupervised text generation // AI Magazine. — 2022. — Т. 43, № 4. — С. 344—352. — DOI: 10.1002/aaai.12068.

35. Neural text generation in regulatory medical writing / C. Meyer [та ін.] // Frontiers in Pharmacology. — 2023. — Т. 14. — DOI: 10.3389/fphar.2023.1086913.

36. NEUROLOGIC AFesque Decoding: Constrained Text Generation with Lookahead Heuristics / X. Lu [та ін.] // NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference. — Association for Computational Linguistics (ACL), 2022. — С. 780—799.

37. Not All Errors Are Equal: Learning Text Generation Metrics using Stratified Error Synthesis / W. Xu [та ін.] // Findings of the Association for Computational Linguistics: EMNLP 2022 / за ред. Y. Goldberg, Z. Kozareva, Y. Zhang. — Association for Computational Linguistics (ACL), 2022. — С. 6588—6603.

38. *OpenAI.* Introducing ChatGPT. — 11.2022. — URL: https://openai.com/blog/chatgpt.

39. *Pautrat-Lertora A.*, *Perez-Lozano R.*, *Ugarte W.* EGAN: Generatives Adversarial Networks for Text Generation with Sentiments // International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K - Proceedings. Т. 1 / за ред. F. Coenen, A. Fred, J. Filipe. — Science, Technology Publications, Lda, 2022. — С. 249—256.

40. *Rao K. Y.*, *Rao K. S.*, *Narayana S. V. S.* Conditional-Aware Sequential Text Generation In Knowledge-Enhanced Conversational Recommendation System // Journal of Theoretical and Applied Information Technology. — 2023. — Т. 101, № 7. — С. 2820—2836.

41. *Seifossadat E.*, *Sameti H.* Improving semantic coverage of data-to-text generation model using dynamic memory networks // Natural Language Engineering. — 2023. — DOI: 10.1017/S1351324923000207.

42. Semi-supervised Adversarial Text Generation based on Seq2Seq models / H. Le [та ін.] // EMNLP 2022 - Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track. — Association for Computational Linguistics (ACL), 2022. — С. 264—272.

43. *Shahriar S.* GAN computers generate arts? A survey on visual arts, music, and literary text generation using generative adversarial network // Displays. — 2022. — Т. 73. — DOI: 10.1016/j.displa.2022.102237.

44. Synthetic Text Generation with Differential Privacy: A Simple and Practical Recipe / X. Yue [та ін.] // Proceedings of the Annual Meeting of the Association for Computational Linguistics. Т. 1. — Association for Computational Linguistics (ACL), 2023. — С. 1321—1342.

45. Text Generation with Diffusion Language Models: A Pre-training Approach with Continuous Paragraph Denoise / Z. Lin [та ін.] // Proceedings of Machine Learning Research. Т. 202 / за ред. A. Krause [та ін.]. — ML Research Press, 2023. — С. 21051—21064.

46. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews / M. J. Page [та ін.] // BMJ. — 2021. — Т. 372. — n71. — DOI: 10.1136/bmj.n71.

47. Towards Table-to-Text Generation with Pretrained Language Model: A Table Structure Understanding and Text Deliberating Approach / M. Chen [та ін.] // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022 / за ред. Y. Goldberg, Z. Kozareva, Y. Zhang. — Association for Computational Linguistics (ACL), 2022. — С. 8199—8210.

48. TrICy: Trigger-Guided Data-to-Text Generation With Intent Aware Attention-Copy / V. Agarwal [та ін.] // IEEE/ACM Transactions on Audio, Speech, and Language Processing. — 2024. — Т. 32. — С. 1173—1184. — DOI: 10.1109/TASLP.2024.3353574.

49. Two-in-One: A Model Hijacking Attack Against Text Generation Models / W. M. Si [та ін.] // 32nd USENIX Security Symposium, USENIX Security 2023. Т. 3. — USENIX Association, 2023. — С. 2223—2240.

50. XWikiGen: Cross-lingual Summarization for Encyclopedic Text Generation in Low Resource Languages / D. Taunk [та ін.] // ACM Web Conference 2023 - Proceedings of the World Wide Web Conference, WWW 2023. — Association for Computing Machinery, Inc, 2023. — С. 1703—1713. — DOI: 10.1145/3543507.3583405.

51. *Yin X.*, *Wan X.* How Do Seq2Seq Models Perform on End-to-End Data-to-Text Generation? // Proceedings of the Annual Meeting of the Association for Computational Linguistics. Т. 1 / за ред. S. Muresan, P. Nakov, A. Villavicencio. — Association for Computational Linguistics (ACL), 2022. — С. 7701—7710.

52. *Слободянюк А. В.* Огляд статей. — 05.2024. — URL: `https://docs.google.com/spreadsheets/d/e/2PACX-1vR6ZUaeeBjVgVl-do6QXm-Pua-HdztOxjC4DUqunrSDZ_-YSRz-Ng9xktYH9b0LDT502SiVy3YePx9F/pubhtml`.

Додаток А
Карта огляду для статті

1. Бібліографічне посилання
2. Тип документу: стаття у журналі або стаття у матеріалах конференції
3. Назва
4. Рік публікації
5. Країни, які представляють автори
6. Мета статті
7. Використані архітектури нейронних мереж
8. Використані метрики якості
9. Характеристики використаних наборів даних
    - ім'я
    - тип даних: речення, абзац, документ, питання-відповідь, не вказаний
    - розмір
    - формат: CSV, JSON, XML, файли, не вказаний
    - за типом розмітки: розмічені дані, нерозмічені дані
    - якість даних: попередньо не опрацьовані (сирі), попередньо опрацьовані
    - за доступністю: загальнодоступний, приватний, не вказаний
    - посилання
10. Розв'язана задача генерації тексту (для чого застосовувалась нейронна мережа)
11. Мова генерації тексту