

ІНТЕЛЕКТУАЛЬНА СИСТЕМА СЕМАНТИЧНОГО АНАЛІЗУ ТА ПОШУКУ У ВІДЕОКОНТЕНТІ НА ОСНОВІ МУЛЬТИМОДАЛЬНОГО ГРАФА ЗНАНЬ

Ватолін Дмитро Дмитрович,
студент групи ІНБ-2-23.40д

Київського столичного університету імені Бориса Грінченка,
науковий керівник – к.т.н., доц. Носенко Тетяна Іванівна

ВСТУП

Актуальність і постановка проблеми

На сьогоднішній день відеоматеріали стали одним із ключових та найпопулярніших джерел інформації: від навчальних сесій і записів конференцій до наукових доповідей та різноманітних брифінгів. Через це здатність ефективно аналізувати та систематизувати інформацію з відео набуває особливого значення. Проте пошук конкретних фактів, визначень чи відповідей на апитання всередині великих відеоархівів досі залишається актуальною проблемою: користувачам доводиться переглядати значні обсяги відео, щоб знайти потрібний фрагмент.

Зі стрімким розвитком великих мовних моделей (BMM, англ. Large Language Models, LLM) з'явилася можливість залучити цю технологію до вилучення інформації з відео та її подальшого аналізу. Проте суттєвою проблемою BMM є обмежене контекстне вікно, яке не дозволяє опрацювати весь необхідний матеріал відео за один раз. Для подолання цього обмеження зазвичай застосовують архітектуру генерації, доповненої пошуком (Retrieval-Augmented Generation, RAG) [1]. Класичний підхід (Vector RAG) полягає в тому, що текст — у нашому випадку транскрипція відео — розбивається на невеликі фрагменти, кожен з яких перетворюється на векторне представлення і зберігається у векторній базі даних. Коли користувач ставить запитання, система знаходить найближчі за змістом фрагменти, після чого передає їх мовній моделі для формування відповіді користувачу.

Проте класичний векторний пошук не є оптимальним для нашого випадку, оскільки відеоматеріал є значно комплекснішою формою інформації, ніж звичайний текст, що спричиняє низку недоліків:

- 1) векторний пошук охоплює лише фрагменти, схожі за змістом на запит, і не відстежує зв'язків між поняттями, які в довгій лекції можуть бути рознесені далеко одне від одного;
- 2) за відсутності у знайденому фрагменті достатньої інформації модель схильна «галюцинувати», тобто генерувати хибну, вигадану інформацію;
- 3) транскрипції аудіоряду недостатньо для повноти контексту, оскільки значна частина відомостей міститься у відеоряді (слайди, схеми, формули); без належного поєднання аудіо- та відеоінформації контекст, доступний моделі, може бути неповним;
- 4) важко чітко визначити часовий проміжок відео (таймкод), з якого було взято кожную частину відповіді.

Щоб подолати ці обмеження, доцільно застосувати гібридний підхід Graph RAG [2; 3], який поєднує векторний пошук із графом знань (англ. knowledge graph) та додатково збагачує граф інформацією із зовнішніх джерел, наприклад Wikidata [4]. Завдяки цьому мовна модель отримує структурований, фактологічно точний і перевірюваний контекст.

МЕТА ДОСЛІДЖЕННЯ

Метою роботи є підвищення точності та фактологічної достовірності відповідей на запитання за відео- та аудіорядом матеріалу, зокрема зменшення кількості галюцинацій BMM, шляхом розроблення моделі мультимодального графа знань, алгоритмів його зовнішнього збагачення та методу гібридного пошуку.

Для досягнення мети вирішуються такі завдання:

- 1) Аналіз сучасних методів вилучення інформації з відеоконтенту та архітектур генерації, доповненої пошуком (векторної, графової, гібридної);
- 2) Розроблення темпоральної моделі мультимодального графа знань;
- 3) Розроблення алгоритмів вилучення фактів з відео, зв'язування сутностей та збагачення графа із зовнішніх джерел (Wikidata);
- 4) Проектування та програмна реалізація конвеєра гібридного векторно-графового пошуку із зазначенням першоджерел;
- 5) Експериментальне дослідження розробленої системи та оцінювання якості генерації відповідей за спеціалізованими метриками RAG.

Дослідницьке питання роботи сформульовано так: чи підвищує використання мультимодального графа знань із часовою прив'язкою якість пошуку та генерації відповідей за відеоконтентом порівняно зі звичайним векторним пошуком? Відповідно, **гіпотеза дослідження** полягає в тому, що додавання графової складової та зовнішнього збагачення збільшує повноту знайденого контексту і зменшує частку необґрунтованих тверджень у відповідях, насамперед для багатокрокових запитань, відповідь на які розподілена між кількома фрагментами відео.

Аналіз останніх досліджень і публікацій

З метою оцінювання потенціалу та практичної цінності системи VerifiedLectureKnowledgeGraph її порівняно з наявними аналогами, які варто поділити на три групи: комерційні рішення для роботи з відеоматеріалами, промислові платформи відеоаналітики та наукові прототипи 2025–2026 років.

Комерційні продукти

Найближчим аналогом системи VerifiedLectureKnowledgeGraph є сервіс **NotebookLM (Google)**. Він здатний опрацьовувати джерела різних типів, зокрема транскрипції відео, з яких отримує інформацію та надалі її систематизує [5]. Перевага сервісу полягає в тому, що він чітко зазначає джерела отриманої інформації та підкріплює відповіді цитатами з них. Також NotebookLM має знижений ризик галюцинацій: за відсутності потрібної інформації в джерелах він прямо повідомляє про це. Проте його слабкість у тому, що відеоряд (слайди, формули, схеми) не аналізується, а опрацьовується лише транскрипція, що робить доступний контекст обмеженим. Знання не

структуруються у вигляді графа, архітектура є закритою та пропрієтарною, а збагачення із зовнішніх баз знань відсутнє.

Ask YouTube (Google) — функція діалогового пошуку та запитань щодо відео на платформі YouTube, доступна передплатникам YouTube Premium; вона дає змогу отримати чіткі відповіді на запитання, пов'язані з конкретним відео, разом із ключовими тезами та таймкодами [6]. Проте обмеженість контексту одним відео, залежність від платформи YouTube (власні відеоархіви не підтримуються) та нерозкриття джерел відповідей роблять цей продукт придатним лише для окремих ситуацій.

Асистенти для онлайн-конференцій, такі як Otter.ai, Fireflies.ai та Fathom, використовуються для транскрибування та узагальнення дискусій під час дзвінків, а також дають змогу ставити запитання за транскриптами власних записів. Ці системи орієнтовані на ділові зустрічі, працюють переважно з текстовою модальністю та не здатні самостійно структурувати знання.

Платформи навчального відео Panopto та Kaltura забезпечують пошук у відеоархіві закладу освіти чи організації за словами з транскрипції та за текстом слайдів, розпізнаним засобами OCR, з переходом до відповідного моменту відео. Проте такий пошук є лише лексичним, через що він не здатний генерувати відповіді та не враховує зв'язків між поняттями.

TwelveLabs — комерційна платформа розуміння відео. Її модель Marengo 3.5 (2026) формує спільний простір вкладень для відео, аудіо, зображень і тексту та підтримує пошук за комбінованими запитами, а модель Pegasus генерує текстові відповіді за відео [7]. Це сильний векторний базовий метод, проте граф знань не є компонентом платформи. Також варто зазначити, що моделі доступні лише через програмний інтерфейс (API), що обмежує їх інтеграцію в інші середовища.

Промислові платформи відеоаналітики

NVIDIA Video Search and Summarization (VSS) Blueprint [8] — еталонна архітектура агентів пошуку, узагальнення та запитань-відповідей за відео: фрагменти відео описуються візуально-мовною моделлю, а модуль контекстно-залежного RAG поєднує векторне та графове сховища, тобто вже реалізує гібридний підхід. Серед усіх розглянутих вище альтернатив цей продукт є найбільш конкурентоспроможним та найближчим до системи VerifiedLectureKnowledgeGraph за архітектурою. Проте через орієнтацію на вузькоспеціалізовані сфери (складські приміщення, міська інфраструктура, безпека) він не пристосований до ширших сфер застосування, зокрема навчального контенту. Також варто зазначити, що платформа потребує спеціалізованого обладнання NVIDIA та ліцензії NVIDIA AI Enterprise і не передбачає збагачення графа із зовнішніх баз знань.

Таким чином, сама концепція поєднання векторного пошуку з графом знань не є новою та вже реалізована в наявних промислових системах. Звідси випливає, що новизну системи VerifiedLectureKnowledgeGraph слід шукати в іншому — у поєднанні властивостей, яких жоден із розглянутих аналогів не має одночасно (див. далі).

Наукові прототипи

У свою чергу, **VideoRAG (Ren et al., 2025) [9]** — двоканална архітектура для наддовгих відео, яка поєднує графове подання текстових знань, зокрема зв'язків між різними відео, з мультимодальним кодуванням контексту. Авторами проєкту оприлюднено набір LongerVideos (понад 160 відео загальною тривалістю понад 134 години) та відкриту реалізацію. Граф не має явної часової прив'язки тверджень до фрагментів, а зовнішнє збагачення не передбачене, що відрізняє цей підхід від системи VerifiedLectureKnowledgeGraph.

Варто зазначити, що інший підхід із такою самою назвою — **VideoRAG (Jeong et al., 2025) [10]** — розв'язує задачу пошуку релевантних відео в корпусі та генерації відповідей з безпосереднім використанням візуальних і текстових ознак. Граф знань у ньому не є центральним компонентом, тому цей підхід доцільно розглядати як базовий метод для порівняння.

ViG-RAG (Cao et al., AAAI 2026) [11], у свою чергу, буде ймовірнісний темпоральний граф знань за транскрипціями, за допомогою якого оцінює достовірність кожного свідчення та виконує адаптивний гібридний пошук за семантичними й часовими ознаками. Ця реалізація є найближчим науковим аналогом за ідеєю темпорального графа.

Graph-to-Frame RAG (Yang et al., CVPR 2026) [12] формує граф, що пов'язує сутності, події, просторові відношення та зовнішні світові знання. Знайдений підграф передається моделі у вигляді додаткового «кадру міркування», залишаючи чіткий слід доказів. Такий принцип підтверджує актуальність перевірюваності відповідей.

Vgent (Shen et al., NeurIPS 2025) [13] за принципом своєї роботи організовує довге відео у граф пов'язаних кліпів, додаючи етап міркування для фільтрації нерелевантних результатів пошуку.

Object-Centric Framework for Video Moment Retrieval (Li et al., AAAI 2026) [14], у свою чергу, використовує графи сцен на рівні об'єктів для локалізації моментів відео. Така система демонструє цінність структурних відношень для пошуку та структурування знань, але не передбачає генерації відповідей.

Окремо варто зазначити графові методи RAG для тексту, а саме GraphRAG [2] та LightRAG [3]. Вони не працюють з відеомодальністю, проте їхні загальні принципи побудови графа засобами BMM і його використання під час пошуку [15,19] є корисними безпосередньо для цієї роботи. Результати порівняння системи VerifiedLectureKnowledgeGraph (далі — VLKG) з аналогами за ключовими критеріями зведено в таблиці 1.

Таблиця 1

Порівняння системи VerifiedLectureKnowledgeGraph з існуючими аналогами за ключовими критеріями

Рішення	Модальності	Граф знань	Таймкоди у відповіді	Зовнішнє збагачення	Посилання на джерела	Доступність, вимоги
NotebookLM (Google)	аудіо / транскрипт	–	частково (цитати на фрагменти)	–	+	закритий сервіс
Ask YouTube (Google)	відео платформи	–	частково	–	–	лише YouTube, Premium
Otter.ai, Fireflies.ai, Fathom	аудіо / транскрипт	–	частково	–	частково	закриті сервіси, власні записи
Panopto, Kaltura	аудіо + OCR слайдів	–	+ (перехід до моменту)	–	н/з (пошук без генерації)	комерційні платформи
TwelveLabs (Marengo, Pegasus)	відео, аудіо, зображення, текст	–	+ (сегменти)	–	–	закриті моделі через API
NVIDIA VSS Blueprint	відео (VLM), аудіо	+	+	–	частково (кліпи)	відкритий код, GPU NVIDIA, ліцензія
VideoRAG (Ren et al.)	відео, аудіо, текст	+	частково	–	частково	відкритий код
ViG-RAG (Cao et al.)	транскрипт, кадри	+ (часовий)	+	–	+ (оцінка достовірності)	дослідницький прототип
Graph-to-Frame RAG (Yang et al.)	відео	+	частково	+ (світові знання)	+ (кадр доказів)	дослідницький прототип
Система VLKG	аудіо, кадри / слайди, текст	+ (часовий, з позначкою походження)	+	+ (Wikidata, позначка походження)	+ (таймкод або зовнішнє джерело)	відкриті компоненти, звичайне обладнання

Опис дослідження, його методів і засобів

У системі VLKG використано низку готових технологій, а саме: автоматичне розпізнавання мовлення (зокрема модель Whisper [16]), виділення ключових кадрів, візуально-мовні моделі та OCR, вилучення засобами ВММ трійок «суб'єкт – предикат – об'єкт» (наприклад, «Ньютон — сформулював — закон тяжіння»), з яких і складається граф, пошук інформації за допомогою векторів, концепцію Graph RAG, а також джерела структурованих фактів, такі як Wikidata [4], для зовнішнього збагачення. Ці технології слугують готовими будівельними блоками.

Власний внесок у роботу полягає в такому:

1) Модель темпорального мультимодального графа знань для аналізу навчального відео, у якій кожна сутність і кожен зв'язок мають часовий

інтервал, модальність-джерело (мовлення, слайд, кадр) та показник достовірності;

2) Механізм збагачення графа із зовнішніх джерел з обов'язковим маркуванням походження (provenance) кожного доданого факту та фільтрацією за релевантністю щодо контексту конкретного відео;

3) Метод гібридного пошуку, результатом якого є не лише відповідь, а й перевірювані посилання — таймкод відео або ідентифікатор зовнішнього факту, на які спиралася модель; при цьому твердження без джерела до відповіді не потрапляють;

4) Орієнтація на навчальний контент (лекції, доповіді, наукові презентації тощо);

5) здатність працювати на звичайному обладнанні без спеціалізованих графічних прискорювачів, на відміну від промислових рішень;

6) експериментальна перевірка внеску саме графової складової за умови однакового тестового корпусу, однакових тестових запитів та однакових базових моделей для всіх конфігурацій.

Конкретні програмні засоби реалізації буде обґрунтовано на етапі проектування. Розглядаються як готові каркаси для застосунків на основі BMM, так і власна реалізація з мінімальною кількістю залежностей. Як сховища розглядаються графова база даних Neo4j та векторні бази Qdrant або Chroma, як система розпізнавання мовлення — Whisper. Важливо зазначити, що остаточний вибір визначатиметься вимогами до відтворюваності результатів та доступних обчислювальних ресурсів.

Головною метою експерименту є перевірка сформульованої гіпотези, тобто кількісне оцінювання внеску графової складової та механізму зовнішнього збагачення в якість відповідей.

Тестовий корпус. Планується сформувати власну базу з 5–10 навчальних відео (лекції та доповіді тривалістю 30–90 хвилин) для тестування системи саме на навчальному контенті, оскільки наявні бенчмарки створені для задач іншого виду: LongerVideos [9] — для довгих відео загального характеру, EASG-Bench [17] — для егоцентричного відео дій. Для тестового набору буде підготовлено від 50 до 100 запитань різних типів з еталонними відповідями та таймкодами. За структурою вони поділятимуться на фактичні, відповідь на які міститься в одному фрагменті; багатокрокові, де система має знайти частини відповіді в різних фрагментах відео для повної та якісної відповіді; такі, що потребують зовнішніх знань; та запитання поза межами контенту — дотичні до теми відео, але без відповіді в ньому — для перевірки стійкості до галюцинацій.

Конфігурації порівняння: (A) векторний RAG за транскрипцією; (B) графовий RAG без векторної складової; (C) гібридний пошук без зовнішнього збагачення; (D) гібридний пошук зі збагаченням із Wikidata, тобто повна конфігурація системи VLKG. Незважаючи на різні підходи, усі конфігурації використовують однаковий корпус, однакові запитання та однакові базові моделі розпізнавання мовлення, вкладень і генерації, що, у свою чергу, дозволяє приписати різницю результатів саме досліджуваним компонентам.

Метрики. Якість генерації відповідей оцінюється за спеціалізованими критеріями RAG, що дають змогу об'єктивно оцінити поточний стан продукту. Зокрема, за допомогою бібліотеки Ragas [18] вимірюються повнота контексту (Context Recall), фактологічна відповідність контексту (Faithfulness) та релевантність відповіді (Answer Relevance). Додатково вимірюються: точність посилань — частка цитованих таймкодів, що потрапляють у вікно ± 15 с від еталонного таймкоду з тестового набору; частка коректних відмов на запитаннях поза межами контенту; точність і повнота вилучених сутностей на розміченій вибірці фрагментів; середня затримка формування відповіді.

Критерій підтвердження гіпотези: гіпотеза вважається підтвердженою, якщо конфігурації C і D статистично значуще перевищують конфігурацію A за повнотою контексту та фактологічною відповідністю на багатокрокових запитаннях без погіршення результатів на фактичних запитаннях, а конфігурація D додатково перевищує конфігурацію C на запитаннях, що потребують зовнішніх знань.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Концепція продукту

Назва системи VerifiedLectureKnowledgeGraph акцентує на перевірюваному графі знань лекційних матеріалів: в її основі лежить головний принцип — перевірюваність кожного твердження у відповіді. Принцип роботи полягає в тому, що користувач завантажує відео або набір лекцій, а система опрацьовує їх у фоновому режимі та надає чат-інтерфейс. Кожна відповідь чат-бота на запитання користувача складається з тексту та переліку джерел двох типів. Перший — фрагмент відео з інтервалом часу (наприклад, [Відео 2, 12:34–13:10]) з можливістю переходу до відповідного моменту для верифікації; другий — зовнішній факт (наприклад, [Wikidata: Q937]), якого у відео немає і який позначається як зовнішнє джерело, що може потребувати додаткової перевірки. Якщо контекст не містить відповіді на поставлене запитання, система чітко повідомляє про відсутність інформації замість надання неправдивої або вигаданої відповіді. Таким чином, продукт поєднує зручність NotebookLM із перевірюваністю відповідей на рівні таймкодів і прозорим розмежуванням знань із відео та доданих ззовні, а також не прив'язує користувача до конкретної платформи, як-от YouTube.

Модель мультимодального графа знань

Граф знань подано у вигляді формули $G = (V, E, \tau, \pi)$, у якій V — множина вузлів, E — множина зв'язків, τ — функція часової прив'язки, що ставить у відповідність кожному елементу графа інтервал $[t_{\text{start}}, t_{\text{end}}]$ у відео, а π — функція походження, яка для кожного елемента визначає його джерело: мовлення, візуальний ряд або зовнішню базу знань.

Попередньо визначено такі типи вузлів: Video (відео), Segment (фрагмент транскрипції з часовими межами), Frame (ключовий кадр з розпізнаним текстом та описом), Entity (сутність: особа, організація, термін, подія), Concept (поняття предметної області) та ExternalFact (факт із зовнішнього джерела з його ідентифікатором). За типами зв'язки поділяються на MENTIONED_IN (сутність згадана у фрагменті або кадрі, з інтервалом часу та модальністю),

RELATED_TO (змістовне відношення між сутностями з типом, показником достовірності та посиланням на фрагмент-джерело), FOLLOWS (часовий порядок фрагментів), SAME_AS (ототожнення сутності з елементом Wikidata) та ENRICHED_WITH (зв'язок сутності із зовнішнім фактом). Модель є попередньою і уточнюватиметься на етапі проєктування.

Часова прив'язка та походження є обов'язковими атрибутами, оскільки саме вони дають можливість формувати посилання у відповідях та відокремлювати знання внутрішнього походження, тобто з відеоряду, від знань із зовнішніх джерел.

Узагальнена архітектура

Узагальнену структурну схему системи VLKG наведено на рисунку 1. Система складається з п'яти функціональних рівнів.

Рівень завантаження та попередньої обробки. З відео відокремлюється аудіодоріжка, яка транскрибується в текст із часовими мітками на рівні фраз. Паралельно із цим процесом виділяються ключові кадри відеоматеріалу (за зміною сцени або з фіксованим кроком), до яких застосовуються технологія OCR та візуально-мовна модель для отримання текстового опису.

Рівень вилучення знань. Велика мовна модель вилучає з транскрипції та описів кадрів сутності й відношення між ними та структурує їх, прив'язуючи кожне до відповідного фрагмента-джерела. Також виконуються дедуплікація та злиття сутностей, отриманих з різних модальностей, для формування цілісної узагальненої картини.

Рівень зберігання. Структуровані знання зберігаються паралельно у графовому сховищі (структура, час, походження) та у векторному сховищі (вкладення фрагментів транскрипції та описів кадрів). Поєднання обох типів сховищ дає змогу використовувати кожне з них за призначенням: граф — для структурних і часових зв'язків, вектори — для пошуку за змістом. Обидва сховища пов'язані спільними ідентифікаторами фрагментів.

Рівень збагачення. Для ключових понять та сутностей з відеоматеріалу із зовнішніх джерел, наприклад Wikidata, добираються кандидати на відповідність. Після цього виконується зняття неоднозначності за контекстом відео, а SPARQL-запитами отримуються пов'язані факти, які фільтруються за релевантністю з подальшим записом у граф і позначкою зовнішнього походження.

Рівень пошуку та генерації. Відповідає за реалізацію гібридного пошуку інформації (описаний нижче) та формування відповідей з посиланнями на джерела.

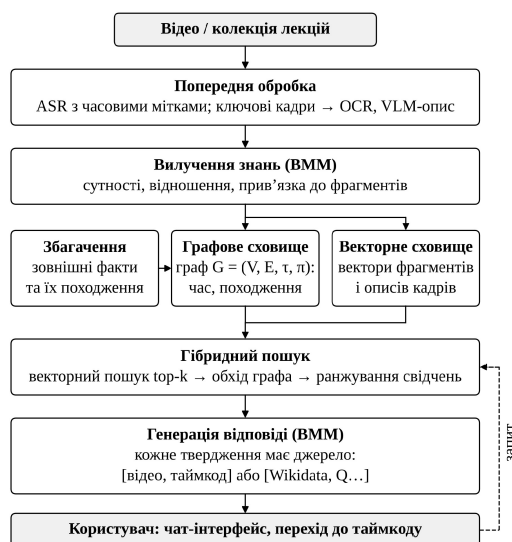


Рис. 1. Узагальнена структурна схема системи VerifiedLectureKnowledgeGraph

Алгоритм гібридного пошуку

Загалом алгоритм складається з таких кроків:

- 1) конвертація запиту користувача у векторну форму;
- 2) векторний пошук k найближчих фрагментів транскрипції та описів кадрів;
- 3) визначення сутностей — точок входу в граф — з тексту запиту та зі знайдених фрагментів;
- 4) сканування графа від точок входу на глибину одного-двох кроків з урахуванням часового порядку фрагментів та збір пов'язаних сутностей, відношень і зовнішніх фактів;
- 5) об'єднання та ранжування зібраних свідчень з урахуванням векторної близькості, зв'язності у графі та достовірності;
- 6) формування контексту для мовної моделі, у якому кожне свідчення супроводжується даними про його походження (метаданими);
- 7) генерація відповіді з обов'язковим цитуванням та перевіркою кожного твердження на наявність джерела; твердження, не підкріплені джерелом, вилучаються або позначаються як непідтверджені.

Функція ранжування, що виконується на кроці 5, та глибина обходу графа є лише параметрами, вплив яких планується дослідити експериментально.

ВИСНОВКИ

Результати проведеного аналізу показали, що системи, які поєднують у собі мультимодальний аналіз відео, графи знань та генерацію, доповнену пошуком, вже існують як у вигляді промислових платформ (NVIDIA VSS, TwelveLabs), так і наукових прототипів (VideoRAG, ViG-RAG, Graph-to-Frame RAG). Через цю концепцію гібридного пошуку як таку не можна вважати новою. Проте варто зазначити, що жодне з розглянутих рішень не поєднує в єдине ціле часову прив'язку знань на рівні фрагментів навчального відео, збагачення із зовнішньої бази знань з явним маркуванням походження, формування перевірюваних посилань у відповідях та відтворюваність на

звичайному обладнанні. Таким чином, хоча окремі технології не є новими, саме їх поєднання в системі VLKG робить її унікальною та вартою уваги.

Отже, у цих тезах запропоновано концепцію системи VLKG, що заповнює відповідну нішу в освітній сфері, попередню модель темпорального мультимодального графа знань, узагальнену архітектуру та алгоритм гібридного пошуку, а також план експерименту з чотирма конфігураціями та набором метрик. Подальша робота передбачає уточнення архітектурних рішень, реалізацію прототипу та експериментальну перевірку сформульованої гіпотези.

ДЖЕРЕЛА

1. Lewis P., Perez E., Piktus A. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*. 2020. Vol. 33. P. 9459–9474.
2. Edge D., Trinh H., Cheng N. et al. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv preprint arXiv:2404.16130*. 2024.
3. Guo Z., Xia L., Yu Y., Ao T., Huang C. LightRAG: Simple and Fast Retrieval-Augmented Generation. *arXiv preprint arXiv:2410.05779*. 2024.
4. Vrandečić D., Krötzsch M. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*. 2014. Vol. 57, No. 10. P. 78–85.
5. NotebookLM adds audio and YouTube support, plus easier sharing of Audio Overviews. Google: The Keyword. 2024. URL: <https://blog.google/technology/ai/notebooklm-audio-video-sources/> (дата звернення: 08.10.2026).
6. Bonggolto J. YouTube is testing a chat-style search that cuts the scrolling. *Android Authority*. 28.04.2026. URL: <https://www.androidauthority.com/youtube-ai-chatbot-search-experiment-3661438/> (дата звернення: 08.10.2026).
7. Marengo 3.5. TwelveLabs Blog. 2026. URL: <https://www.twelvelabs.io/blog/marengo-3.5> (дата звернення: 08.10.2026).
8. NVIDIA Blueprint for Video Search and Summarization (VSS): documentation, version 3.0.0. NVIDIA. 2026. URL: <https://docs.nvidia.com/vss/3.0.0/overview/latest/index.html> (дата звернення: 08.10.2026).
9. Ren X., Xu L., Xia L., Wang S., Yin D., Huang C. VideoRAG: Retrieval-Augmented Generation with Extreme Long-Context Videos. *arXiv preprint arXiv:2502.01549*. 2025.
10. Jeong S., Kim K., Baek J., Hwang S. J. VideoRAG: Retrieval-Augmented Generation over Video Corpus. *Findings of the Association for Computational Linguistics: ACL 2025*. 2025.
11. Cao Z., Liu A., He Y., Li J., Zhang B., Wang Z. ViG-RAG: Video-aware Graph Retrieval-Augmented Generation via Temporal and Semantic Hybrid Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2026. Vol. 40, No. 1. P. 48–56. DOI: 10.1609/aaai.v40i1.36963.
12. Yang S., Yu W., Liu Z. et al. Graph-to-Frame RAG: Visual-Space Knowledge Fusion for Training-Free and Auditable Video Reasoning. *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2026. arXiv:2604.04372.

13. Shen X., Zhang W., Chen J., Elhoseiny M. Vgent: Graph-based Retrieval-Reasoning-Augmented Generation for Long Video Understanding. Advances in Neural Information Processing Systems. 2025. Vol. 38.

14. Li Z., Wong Y., Yamazaki S., Liu J., Kankanhalli M. Object-Centric Framework for Video Moment Retrieval. Proceedings of the AAAI Conference on Artificial Intelligence. 2026. arXiv:2512.18448.

15. Peng B., Zhu Y., Liu Y. et al. Graph Retrieval-Augmented Generation: A Survey. arXiv preprint arXiv:2408.08921. 2024.

16. Radford A., Kim J. W., Xu T. et al. Robust Speech Recognition via Large-Scale Weak Supervision. Proceedings of the 40th International Conference on Machine Learning (ICML). PMLR. 2023. Vol. 202. P. 28492–28518.

17. Rodin I. et al. EASG-Bench: Video Q&A Benchmark with Egocentric Action Scene Graphs. arXiv preprint arXiv:2506.05787. 2025.

18. Es S., James J., Espinosa-Anke L., Schockaert S. RAGAs: Automated Evaluation of Retrieval Augmented Generation. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. 2024. P. 150–158.

19. Abramov, V., Astafieva, M., Boiko, M., Bodnenko, D., Bushma, A., Vember, V., Hlushak, O., Zhylytsov, O., Ilich, L., Kobets, N., Kovaliuk, T., Kuchakovska, H., Lytvyn, O., Lytvyn, P., Mashkina, I., Morze, N., Nosenko, T., Proshkin, V., Radchenko, S., & Yaskevych, V. (2021). Theoretical and practical aspects of the use of mathematical methods and information technology in education and science.

URL: <https://doi.org/10.28925/9720213284km>