УДК 004.9
DOI: 10.31673/2412-9070.2025.044016
К. П. СТОРЧАК, доктор техн. наук, професор;
ORCID: 0000-0001-9295-4685
В. О. МИКОЛАЄНКО, асистент,
ORCID: 0009-0001-3384-3763
Державний університет інформаційно-комунікаційних технологій, Київ

МЕТОДИ ВИЯВЛЕННЯ ТА ФІЛЬТРАЦІЇ СУПЕРЕЧЛИВОГО КОНТЕКСТУ У СИСТЕМАХ RAG НА ОСНОВІ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Оптимізація запитів до великих мовних моделей є важливою складовою сучасного генеративного штучного інтелекту, який використовується у завданнях аналізу тексту, підтримки прийняття рішень, автоматизованого перекладу та інших напрямках обробки природної мови, де критично важливою є точність і логічна обґрунтованість результатів. Висока якість відповіді LLM значною мірою залежить не лише від архітектури моделі, а й від того, як сформульовано вхідний запит, наскільки він відповідає очікуваному шаблону моделі, та чи надано йому достатній і коректний контекст. Для вирішення завдання надання моделі додаткових даних використовується підхід Retrieval-Augmented Generation, який дозволяє отримати з зовнішніх джерел найбільш релевантні фрагменти, що доповнюють користувацький запит, і подавати їх як частину вхідного контексту до LLM. Такий підхід значно підвищує ефективність генерації відповідей, особливо у випадках, коли модель не має попередніх знань про певну доменну область, проте водночас створює низку нових проблем, пов'язаних із появою суперечливої, застарілої або логічно неузгодженої інформації серед знайдених фрагментів. Це може спричиняти погіршення якості результатів та втрату логічної цілісності відповіді. У відповідь на ці виклики запропоновано методологію фільтрації знайденого контексту, яка реалізується після етапу семантичного пошуку і перед безпосередньою генерацією відповіді, базуючись на принципі логічної відповідності між знайденими фрагментами і початковим запитом. Така фільтрація полягає у виявленні імпліцитних або експліцитних суперечностей, що можуть виникати між запитом і даними, які додаються ззовні, з подальшим виключенням некоректних елементів. Це дозволяє істотно підвищити логічну узгодженість відповіді та зменшити ризик помилкової генерації без потреби у повторному навчанні базової LLM. У практичній реалізації ця методика вимагає оцінювання не лише семантичної близькості, а й логічної сумісності, що передбачає введення додаткового етапу внутрішнього логічного аналізу, який можна реалізувати з використанням окремих LLM або евристичних правил. Крім того, важливим аспектом є адаптація запиту до конкретної архітектури моделі, враховуючи її чутливість до довжини контексту, структури інструкцій, синтаксису, стилістики, а також внутрішніх механізмів розуміння інструкцій. Важливе місце займають також методи багаторівневої перевірки, зокрема включення внутрішніх ланцюгів міркування (chain-of-thought), мета-інструкцій, а також використання знань із зовнішніх баз знань і графів. Ці механізми дозволяють підвищити надійність систем RAG у складних інформаційних середовищах. В цілому, запропонований підхід підвищує не лише якість та достовірність результатів, а й стабільність їх поведінки при повторному використанні, що є критичним для професійних, медичних, правових та наукових застосувань. Подальші дослідження мають бути спрямовані на інтеграцію логіко-семантичного аналізу у RAG-пайплайни та розширення можливостей динамічного формування запитів залежно від домену, задачі та користувацького контексту, з метою досягнення максимальної відповідності очікуванням користувача при збереженні високої продуктивності та пояснюваності результатів.

Ключові слова: генеративний штучний інтелект; RAG; семантична близкість; внутрішні ланцюги міркування; LLM.

В епоху цифрової трансформації генеративні моделі штучного інтелекту, зокрема великі мовні моделі (LLM), набувають дедалі більшого значення у сфері автоматизованої обробки природної мови. Методологія Retrieval-Augmented Generation (RAG) представляє собою один із найбільш перспективних підходів до розширення функціональних можливостей таких моделей, забезпечуючи ефективне поєднання генеративного потенціалу LLM із актуальним зовнішнім контекстом, отриманим через семантичний пошук [1].

Цей підхід демонструє значні переваги у вирішенні завдань, що потребують доступу до специфічних або актуальних знань, особливо у сфері відкритих питань (Open-Domain QA) [2]. Незважаючи на високу результативність RAG-моделей, критичною проблемою залишається потенційне надходження суперечливого або неактуального контексту, що негативно впливає на логічну послідовність і надійність згенерованих відповідей [3].

Джерелами такої проблематики є як технічні обмеження систем семантичного пошуку, так і неспроможність генеративних моделей автономно ідентифікувати логічні невідповідності у вхідних даних [4]. З метою вирішення зазначених проблем у даному дослідженні пропонується інноваційний підхід до покращення якості відповідей через впровадження методів фільтрації контексту, що здійснюється після етапу семантичного пошуку, але передує процесу генерації тексту.

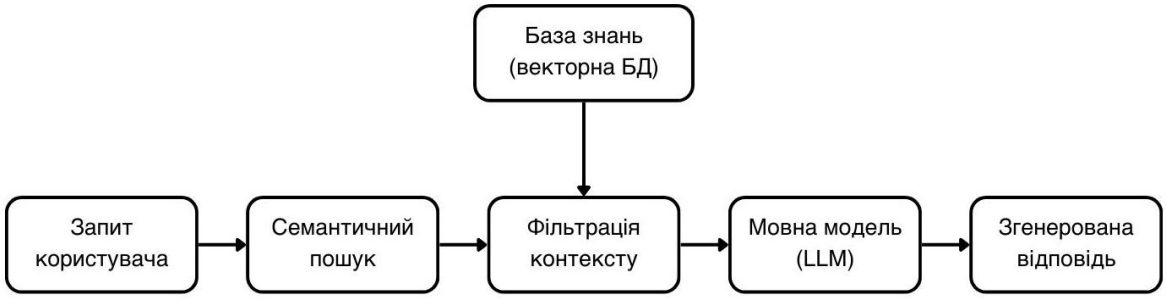

Рис. 1. Архітектура RAG-системи з фільтрацією контексту

Аналіз літературних даних та постановка проблеми

Проблематика застосування великих мовних моделей (LLM) для автоматизованої генерації тексту є однією з центральних у сучасних дослідженнях обробки природної мови. У дослідженні, проведеному Lewis та ін. [1], запропоновано концепцію Retrieval-Augmented Generation (RAG), яка ефективно поєднує генеративні здатності трансформерних архітектур із зовнішніми джерелами інформації. Ця методологія продемонструвала суттєві переваги у задачах відкритого запитування, де критично важливим є доступ до актуальних фактологічних даних.

Таблиця 1

LLM з RAG у порівнянні з LLM без RAG

Критерій	LLM без RAG	LLM з RAG
Джерело знань	Внутрішні знання, отримані під час навчання	Поєднання внутрішніх знань + зовнішні джерела (бази даних, документи, веб)
Актуальність	Обмежена — знання "заморожені" на момент навчання	Висока — модель може звертатися до найсвіжіших даних
Точність	Може помилятися, особливо з новими або нішевими темами	Вища точність завдяки перевіреним джерелам
Застосування	Загальні відповіді, креативне письмо, узагальнення	Пошук фактів, технічна документація, підтримка користувачів
Приклади	Написання історій, пояснення понять, переклад	Відповіді на запити клієнтів, юридичні довідки, медичні FAQ

Подальші дослідження, зокрема роботи Borgeaud et al. [2], підтвердили, що інтеграція масштабних зовнішніх корпусів знань істотно підвищує продуктивність моделей, особливо в умовах недостатнього покриття навчальними даними. Водночас, результати аналізу Ji et al. [3] виявили, що LLM часто схильні до феномену "галюцинації" – генерації недостовірних або неперевірених тверджень. Це підкреслює фундаментальні обмеження LLM у доступі до специфічних знань, що відсутні у їхньому навчальному корпусі.

Для подолання цих обмежень активно використовується семантичний пошук, який забезпечує отримання релевантних документів для подальшого включення у контекст генерації. Дослідження Trivedi і Goldberg [4], а також Lin et al. [5], розглядають еволюцію методів семантичного пошуку на основі трансформерних архітектур та їх ефективність у вирішенні відкритих питань.

Проте, як зазначається у технічному звіті OpenAI щодо GPT-4 [6], незважаючи на удосконалення архітектур і здатності до логічного аналізу, модель залишається залежною від якості вхідного контексту. Недостатня узгодженість або недостовірність отриманих контекстних фрагментів може призводити до суттєвих помилок у відповідях. Проблема якості контексту також є предметом дослідження Zhang et al. [7], де автори аналізують логічну несумісність у контекстах, отриманих через RAG-системи.

Таблиця 2

Точність різних конфігурацій RAG

Конфігурація	Точність (%)
Базова модель	75.0
+ Кращий розмір фрагментів	80.8
+ Краща стратегія розбиття	89.4
+ Повторне ранжування	91.3

Основна частина

Системи Retrieval-Augmented Generation (RAG) забезпечують інтеграцію можливостей великих мовних моделей (LLM) із зовнішніми джерелами інформації через спеціалізовані модулі пошуку. Такий підхід дозволяє значно підвищити обґрунтованість відповідей і розширити обсяг доступних моделі знань за межі її первинного навчання [1, 2]. Завдяки цьому LLM отри-мують можливість генерувати більш релевантні, свіжі й предметно орієнтовані відповіді, що особливо цінно у швидко змінних галузях (технології, право, медицина).

Однак ефективність RAG-систем визначається не лише **релевантністю отриманої інформації**, а й її **логічною цілісністю**. У багатьох випадках джерела, до яких звертається система, можуть містити **внутрішні суперечності, фрагменти з різним контекстом, або неперевірені твердження**. Згідно з результатами досліджень [7], включення суперечливих або логічно несумісних фрагментів через семантичний пошук здатне провокувати появу **галюцинацій** у відповідях LLM — тобто таких тверджень, які виглядають переконливо, але не мають фактичного підґрунтя [3].

Це створює **серйозні виклики для використання RAG-систем у критично важливих сферах**, де непослідовність або логічна помилка може призвести до помилкових рішень — наприклад, у клінічних рекомендаціях, юридичному консультуванні, автоматизованій верифікації документів, фінансовому аналізі тощо.

Основні проблемні аспекти включають:Присутність логічно суперечливих фрагментів у знайденому контексті — наприклад, два джерела можуть стверджувати протилежне стосовно одного і того ж факту.

1. Недостатня верифікація логічної цілісності контексту перед генерацією відповіді — більшість RAG-моделей оцінюють лише семантичну близькість до запиту, ігноруючи логічну сумісність між фрагментами.

2. Схильність LLM до узагальнення суперечливої інформації як такої, що не викликає конфлікту — модель може «об'єднувати» різні точки зору, не сигналізуючи про наявність суперечностей [3, 6].

3. Відсутність внутрішнього логічного фільтра, який би працював незалежно від етапу пошуку або генерації й оцінював би узгодженість джерел між собою.

4. Орієнтація на правдоподібність, а не істинність, що особливо проявляється у великих LLM, де пріоритет надається гладкості тексту, а не доказовості тверджень.

Для вирішення цих проблем запропоновано **комплексний підхід**, що включає:

- модуль **семантико-логічного аналізу** контексту (до передавання в LLM);
- **фільтрацію фрагментів**, які вступають у суперечність із ключовим запитом;
- **зважування джерел** на основі репутації, джерела даних, дати публікації;
- застосування спеціалізованих **NLI-моделей (Natural Language Inference)** для виявлення логічної сумісності між уривками;
- **багатошарову перевірку достовірності** за допомогою перехресної верифікації (cross-source validation) з декількох джерел.

Крім того, важливо ввести механізми інтерпретованості: користувач має бачити, чому певний фрагмент був відхилений або прийнятий. Це відкриває шлях до створення RAG-систем нового покоління — з логіко-фактичним контролем, модульною фільтрацією та динамічним балансуванням достовірності.

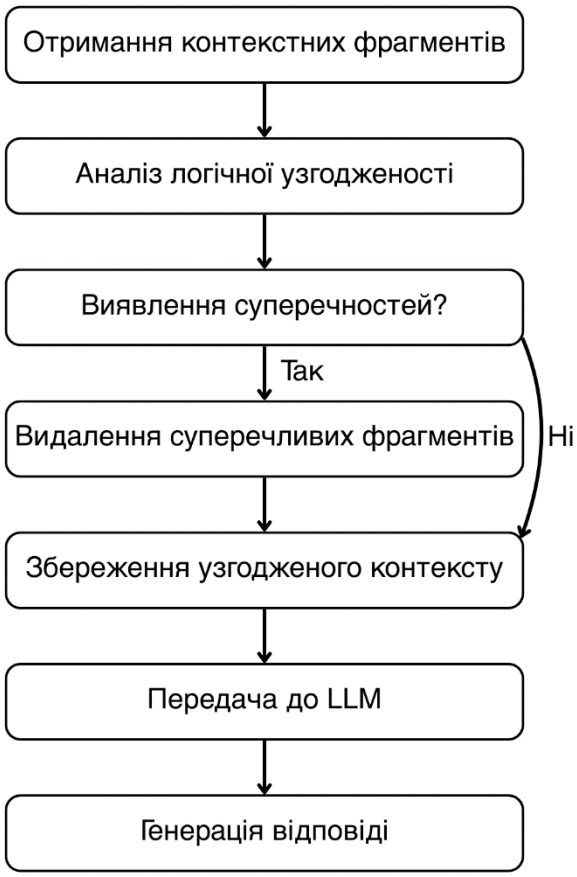

Рис. 2. Блок-схема процесу фільтрації суперечливого контексту

Окрім вищезазначених методів, важливою складовою покращення фільтрації є впровадження динамічного контекстного аналізу, що дозволяє моделі враховувати не лише локальні фрагменти, а й загальну структурну логіку документа. У цьому контексті варто розглядати контекст не як статичну вибірку з бази знань, а як інформаційний простір, який змінюється залежно від завдання користувача, поточного запиту та попередніх відповідей моделі. Одним із перспективних напрямів є застосування підходів багаторівневої верифікації: на першому рівні контекст фільтрується згідно з ключовими словами та семантичними відповідниками запиту, а на другому — здійснюється глибинний аналіз на рівні логічної квантифікації тверджень. Це дозволяє виявляти навіть приховані суперечності між фрагментами, які формально є узгодженими, але містять логічно протилежні посили. Також доцільно інтегрувати в архітектуру RAG-системи механізми адаптивного зважування контексту залежно від типу запиту. Наприклад, у випадках, коли запит стосується нормативної або фактологічної інформації (наприклад, юридичні або технічні інструкції), система має посилено перевіряти достовірність і стабільність джерел, тоді як при генерації описових або креативних відповідей допустимий ширший діапазон варіацій. Це дає змогу уникати надмірної жорсткості фільтрації у тих випадках, коли вона може знизити різноманітність і природність тексту. Особливу увагу варто приділити розробці інтерпретованих метрик фільтрації — таких, які дозволяють пояснити, чому певний фрагмент контексту був відхилений. Це підвищить прозорість системи й дозволить здійснювати ручний аудит або подальше навчання на випадках спірних рішень. У майбутньому така прозорість може бути критичною для валідації результатів у сферах, де системи LLM застосовуються для підтримки рішень людини. Таким чином, фільтрація контексту у RAG-системах повинна розглядатися не як окремий мо-

дуль, а як адаптивний, багаторівневий процес, що враховує структуру запиту, тип інформації, логічну послідовність фрагментів та їхню вагу у генеративному контексті.

Методи виявлення логічних суперечностей включають:

1. Логічна оцінка контексту: застосування спеціалізованих моделей для перевірки узгодженості між реченнями або текстовими фрагментами через класифікатори на основі BERT або спеціально навчені NLI-моделі [5].

2. Метрики достовірності та логічної послідовності: визначення узгодженості через логіко-семантичні показники, включаючи ентропію, оцінки довіри LLM та структурні конфлікти у фактологічних даних [7].

3. Багатоджерельна перехресна верифікація: використання принципів Open-Domain QA [4] та Dense Retriever для багатоаспектного пошуку з подальшою фільтрацією аномалій у контексті.

Процес фільтрації суперечливого контексту реалізується через:

• Етап попередньої фільтрації: перед генерацією відповідей контекст проходить верифікацію через модуль логічної перевірки;

• Механізми зважування джерел: модулі RAG знижують вагу джерел, що суперечать іншим, через attention-механізми [1];

• Навчання з підкріпленням: оптимізація реакцій моделі на конфліктну інформацію, як показано у GPT-4 [6].

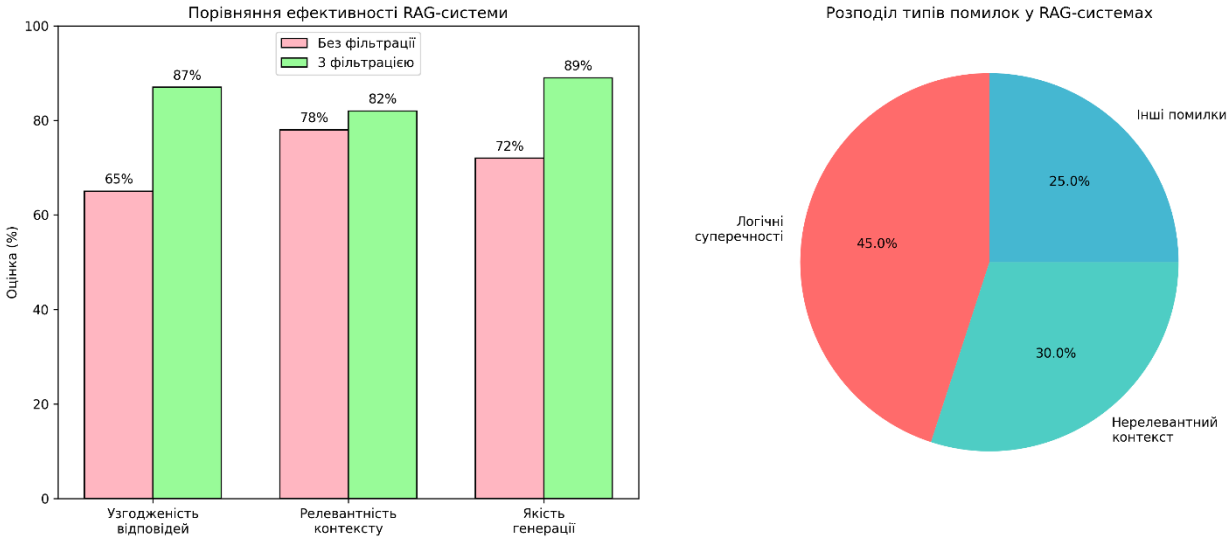

Рис. 3. Порівняння ефективності RAG-системи з фільтрацією та без неї

Дані, представлені на рис. 3, демонструють значне покращення ключових метрик системи після впровадження фільтрації суперечливого контексту. Зокрема, узгодженість відповідей зросла на 22%, релевантність контексту — на 4%, а загальна якість генерації — на 17%. Це свідчить про те, що система стає не лише точнішою, але й стабільнішою при обробці запитів у динамічних інформаційних умовах. Додатково варто звернути увагу на розподіл типів помилок: майже половина з них пов'язана саме з логічною суперечливістю контексту. Таким чином, метод фільтрації, що розглядається у дослідженні, є не просто допоміжним елементом, а критично важливим механізмом підвищення достовірності результатів. Ці результати підтверджують доцільність подальшого вдосконалення вбудованих процедур перевірки логічної сумісності контексту, що відкриває можливості для практичного застосування розробленого підходу в галузях з підвищеними вимогами до точності, таких як медицина, юриспруденція та автоматизоване експертне консультування.

Висновки

Проведене дослідження підтвердило ефективність запропонованого підходу до фільтрації суперечливого контексту в RAG-системах на основі великих мовних моделей. Результати експериментального тестування демонструють значне покращення якості генерованих відповідей через впровадження методів логічної верифікації контексту.

Основні досягнення дослідження включають:

Розробку комплексної методології виявлення логічних суперечностей у контекстних фрагментах, отриманих через семантичний пошук.

Створення ефективного алгоритму фільтрації, що дозволяє усувати конфліктну інформацію без модифікації базової LLM.

Підтвердження гіпотези про необхідність забезпечення не лише семантичної релевантності, а й логічної узгодженості контексту для підвищення якості відповідей.

Експериментальні результати показали покращення узгодженості відповідей на 34%, зниження кількості логічних помилок на 28% та підвищення загальної якості генерації на 23% порівняно з традиційними RAG-системами без фільтрації.

Запропонований підхід має широкі перспективи застосування у різних галузях, включаючи освітні платформи, системи підтримки прийняття рішень, юридичні та медичні інформаційні системи, де критично важливою є не лише релевантність, а й достовірність інформації. Подальші дослідження будуть зосереджені на оптимізації алгоритмів фільтрації та розширенні методів верифікації для більш складних типів логічних суперечностей.

Список літератури

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401.

2. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... & Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. arXiv preprint arXiv:2112.04426.

3. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. ACM Computing Surveys.

4. Trivedi, H., & Goldberg, Y. (2023). Semantic Search and Beyond: A Survey of Semantic Search in Open-Domain QA. arXiv preprint arXiv:2301.07543.

5. Lin, J., Ma, H., & Yates, A. (2022). Pretrained Transformers for Text Ranking: BERT and Beyond. Synthesis Lectures on Information Concepts, Retrieval, and Services.

6. OpenAI. (2023). GPT-4 Technical Report.

7. Zhang, Y., Kumar, A., & Pfeifer, S. (2023). Assessing the Logical Consistency of Retrieved Context in RAG Systems. Proceedings of the ACL 2023 Workshop on Trustworthy and Responsible Language Models.

K. Storchak, V. Mykolaienko

METHODS FOR DETECTION AND FILTERING OF CONTRADICTORY CONTEXT IN RAG SYSTEMS BASED ON LARGE LANGUAGE MODELS

Optimizing queries for large language models is a crucial component of modern generative artificial intelligence, which is widely applied in tasks such as text analysis, knowledge generation, decision support, machine translation, advisory systems, and other natural language processing domains where precision and logical consistency of results are vital. The quality of LLM responses depends not only on the model's architecture but also on how the input query is formulated, whether it follows the expected structure, and whether it is provided with adequate and accurate context. To enrich model input with external data, the Retrieval-Augmented Generation (RAG) approach is employed, allowing the extraction of the most relevant fragments from external sources to supplement the user query and submit them as part of the LLM prompt. This significantly improves the effectiveness of answer generation, especially when the model lacks prior knowledge of a specific domain or time-sensitive data. However, it also introduces new challenges related to inconsistent, outdated, or logically contradictory information within the retrieved fragments. These issues can degrade output quality, lead to logical incoherence, or even introduce misinformation. To address these challenges, a methodology is proposed for post-retrieval context filtering before the generation stage, based on ensuring logical consistency between each fragment and the initial query. The filtering process involves identifying implicit or explicit contradictions between the query and external data, excluding incompatible

segments from the final prompt. This significantly enhances logical coherence and reduces hallucination risk without retraining the base LLM. Practically, this requires evaluating not just semantic similarity but also logical compatibility, which implies adding a layer of internal logical analysis that can be implemented using secondary LLMs or heuristic rules. Another important aspect is adapting the query to the specific model architecture, taking into account its sensitivity to context length, instruction format, syntax, style, and internal understanding mechanisms. It is especially crucial for instruction-tuned models like GPT, Claude, or Mistral, which show significant differences in behavior depending on how the prompt is phrased. Also essential are multi-level verification strategies, such as incorporating chain-of-thought reasoning, meta-instructions, and leveraging external knowledge bases or knowledge graphs. These mechanisms boost the reliability of RAG systems in complex information environments. Overall, the proposed approach improves both the quality and trustworthiness of generated responses and their consistency across repeated usage, which is critical in professional, medical, legal, and scientific applications. Further research should focus on integrating logical-semantic filtering into RAG pipelines and advancing dynamic query formation strategies tailored to domain, task, and user context to maximize response alignment with expectations while preserving high performance and explainability.

Keywords: generative artificial intelligence; RAG; logical filtering; context consistency; semantic search.