УДК 004.8

Д.І. ЮВЖЕНКО
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна,
e-mail: *d.yuvzhenko@kpi.ua*.

С.Г. СТІРЕНКО
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна,
e-mail: *s.stirenko@kpi.ua*.

ПОРІВНЯЛЬНЕ ОЦІНЮВАННЯ СТРАТЕГІЙ СЕГМЕНТАЦІЇ ДОКУМЕНТІВ У СИСТЕМАХ ГЕНЕРАЦІЇ, ДОПОВНЕНОЇ ПОШУКОМ

Анотація. Наведено емпіричне порівняльне дослідження чотирьох стратегій сегментації: фіксованих вікон розміром 256, 512 і 1024 токенів, а також семантичної сегментації на основі великої мовної моделі. Проведено експерименти на довгих змістовно-зв'язних текстах набору даних SQuALITY. Виконано оцінювання на 225 парах запитання–відповідь із використанням метрик Precision@5, Recall@5 (метрики для топ-5 результатів пошуку), якості відповіді (Exact Match, токен-рівневий F1) і середньої затримки пошуку. Отримано результати, що виявляють чіткий компроміс між точністю та повнотою пошуку, зумовлений гранулярністю: менші фрагменти забезпечують вищу точність, тоді як більші суттєво підвищують повноту та покращують якість відповідей за F1-метрикою. У межах цього експериментального дослідження, семантична сегментація демонструє конкурентні результати, але не показує стабільної переваги порівняно з фіксованими вікнами розміром 512–1024 токенів. Зафіксовано зниження затримки пошуку під час використання більших сегментів, що пояснюється меншою щільністю векторного індексу. Запропоновано відтворювану процедуру оцінювання та практичні рекомендації щодо вибору стратегії сегментації для ефективних RAG-систем.

Ключові слова: доповнена пошуком генерація, RAG, сегментація тексту, семантичний пошук, довгі документи, стратегії сегментації.

DOI 10.34229/KCA2522-9664.26.3.4

ВСТУП

Генерація, доповнена пошуком (retrieval-augmented generation, RAG) [1], стала провідною парадигмою для розширення можливостей великих мовних моделей (ВММ) за рахунок зовнішніх джерел знань, забезпечуючи підвищену фактологічну обґрунтованість, інтерпретованість та стійкість у широкому спектрі задач, що потребують інтенсивної роботи з інформацією [2]. Поєднання модуля пошуку, який відбирає релевантні фрагменти документа, з генеративною моделлю, що формує відповідь на основі цих фрагментів, дає змогу RAG-системам долати обмеження фіксованих параметрів моделей та оперативно оновлювати знання без повторного навчання. Завдяки таким перевагам RAG добре зарекомендував себе в задачах, де корисна інформація може бути розподілена по великому змістовно-зв'язному або технічному тексту.

У цій роботі термін «фрагмент» (інколи «сегмент») позначає частину документа, отриману внаслідок сегментації.

Одним із ключових, але часто недооцінених компонентів RAG-систем, є сегментація (chunking) — процес розбиття документів на менші фрагменти, придатні для пошуку та індексації [3]. Сегментація визначає гранулярність пошукового простору, семантичну цілісність отриманого контексту та структуру вхідних даних для генеративної моделі. Попри важливість, оптимальні стратегії сегментації досі не стандартизовані, а кількісні порівняння різних розмірів фрагментів чи методів сегментації описано в працях лише частково. Переважна більшість сучасних RAG-систем використовують фіксовані вікна токенів за замовчуванням, тоді