https://doi.org/10.31891/2219-9365-2026-87-6
УДК 004.[05+94]

МОЧЕРНЮК Юрій
Національний лісотехнічний університет України
https://orcid.org/0009-0002-1370-363X
e-mail: mocherniuk.iurii@nltu.lviv.ua
МОКРИЦЬКИЙ Андрій
Національний лісотехнічний університет України
https://orcid.org/0009-0009-4744-8981
e-mail: andriy.mokrytskyy@nltu.edu.ua

СЕСІЙНА ПАМ'ЯТЬ ТА СТИСКАННЯ КОНТЕКСТУ ДЛЯ ГЕНЕРУВАННЯ ВІДПОВІДЕЙ У МЕСЕНДЖЕРАХ З НИЗЬКИМИ ЗАТРИМКАМИ

У роботі запропоновано швидку архітектуру для фактичної генерації відповідей у месенджерах — сесійна пам'ять із керованим стисканням, гібридний (BM25 + вектори) пошук із повторним ранжуванням та оптимізації інференсу. Експерименти на відкритих і анонімізованих діалогах показали збереження/покращення якості (nDCG@k, Recall@k, людські й автоматизовані оцінки), зниження затримки й внесок компонентів за результатами абляції.

Ключові слова: великі мовні моделі, месенджери, сесійна пам'ять, стиснення контексту, гібридний пошук, повторне ранжування.

MOCHERNIUK Yurii, MOKRYTSKYI Andrii
Ukrainian National Forestry University

SESSION MEMORY AND CONTEXT COMPRESSION FOR LOW-LATENCY RESPONSE GENERATION IN MESSAGING APPLICATIONS

The generation of responses in messaging applications under long dialogs and strict response-time requirements is analyzed. An architecture is proposed in which session memory forms a compact, incrementally updated representation of the conversation with citation anchors; context compression selects salient units under a fixed token budget; hybrid retrieval (BM25 and vector representations with messenger-specific signals) followed by reranking refines candidates; and the generation module provides streaming output with inference accelerations (efficient attention, speculative decoding, weight quantization, key–value caching). A methodology for factuality control with automated consistency checking and standardized inline citation is developed. Experimental evaluation is conducted on public retrieval-augmented and dialogue corpora complemented with anonymized fragments of real conversations; retrieval quality (nDCG@k, Recall@k), response usefulness and factuality (human side-by-side judgments and automatic metrics), as well as timing characteristics (median and 95th-percentile end-to-end latency with stage breakdown) are reported. The impact of context-compression degree and of hybrid-retrieval and reranking parameters on source consistency is investigated, with comparisons to baselines without session memory, without compression, and without reranking. Latency reductions are demonstrated while preserving or improving quality; the contribution of individual components is established via ablation, and practical recommendations are provided for different response-time budgets (mobile and server settings). The scientific novelty lies in unifying session memory with controlled context compression, hybrid retrieval, and reranking within a single system augmented with engineering accelerations to ensure low latency without loss of factuality; the practical significance is a robust methodology suitable for real-world deployment under privacy and reproducibility requirements.

Keywords: large language models, messaging applications, session memory, context compression, hybrid retrieval, reranking.

Стаття надійшла до редакції / Received 10.05.2026
Прийнята до друку / Accepted 25.06.2026
Опубліковано / Published 10.09.2026

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ'ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

Інтенсивне впровадження помічників на основі великих мовних моделей у месенджерах супроводжується жорсткими вимогами до часу відповіді та збереження релевантності в довготривалих багатогілкових бесідах. Просте збільшення обсягу вхідного контексту не гарантує відбору найсуттєвіших фрагментів і призводить до зростання затримок, зокрема через відому деградацію корисності довгих контекстів [1–3, 7]. На релевантність і стійкість відповіді додатково впливають характерні для месенджерів сигнали — відповіді «ниткою», реакції, пересилання, вкладення та голосові повідомлення — що ускладнює пошук потрібних фрагментів і вимагає поєднання **лексичних методів (BM25)** та **векторних представлень** з подальшим повторним ранжуванням [1–4]. Водночас практичні сценарії накладають обмеження на обчислювальні ресурси та пам'ять, тож потрібні інженерні засоби скорочення часу оброблення (ефективні реалізації уваги, прогнозне декодування, квантування, керування кешем ключ–значення) [10–15].

У зв'язку з цими виникає необхідність у розгляді підходу, який поєднує сесійну пам'ять як компактне

інкрементально оновлюване подання діалогу з керованим стисненням контексту та гібридним пошуком із повторним ранжуванням. Формування відповіді може виконуватись з використанням засобів прискорення інференсу та з контролем фактичної коректності через цитатні покажчики та автоматизовані метрики узгодженості [1–4, 8, 10–13, 18, 19]. Для швидкого відбору кандидатів доцільно застосувати індексні структури наближеного пошуку найближчих сусідів (FAISS, HNSW), що забезпечують ефективність на великих колекціях [5, 6], а стислий контекст формується з орієнтацією на збереження ключових одиниць під фіксований бюджет токенів [7, 8].

АНАЛІЗ ОСТАННІХ ДОСЛІДЖЕНЬ ТА ПУБЛІКАЦІЙ.

Підходи генерації, доповненої пошуком, набули статусу базової парадигми для знаннєво-інтенсивних завдань: у роботі [1] запропоновано інтеграцію зовнішніх джерел знань із мовною моделлю для зменшення галюцинацій і підвищення точності відповіді. Модель попереднього навчання з пошуком REALM [2] демонструє, що явне включення пошукового компонента у представлення запиту покращує отримання релевантних фрагментів. Підхід Fusion-in-Decoder [3] показує ефективність багатопрохідного збирання пасажів із подальшою обробкою їх у декодері. Водночас ці рішення головно працюють зі статичними запитами, тоді як месенджерні діалоги мають динамічну, багатониткову структуру та суворі вимоги до часу відповіді.

Ефективність відбору релевантних фрагментів спирається на поєднання лексичних методів і векторних представлень: автори дослідження [4] пропонують ColBERTv2 з «пізньою взаємодією», який забезпечує високу точність завдяки токен-рівневій відповідності за помірних витрат у індексі. Для швидкої обробки великих колекцій застосовуються програмні засоби наближеного пошуку найближчих сусідів на GPU, зокрема FAISS [5], а також графові структури HNSW [6], що забезпечують малі затримки при масштабуванні. Разом із тим у літературі бракує робіт, де вагування характерних для месенджерів сигналів — відповідей «ниткою», реакцій, пересилань — інтегровано безпосередньо у функції відбору та повторного ранжування.

Проблему зниження корисності довгих підказок у мовних моделях системно окреслено в дослідженні «Lost in the Middle» [7], де показано деградацію впливу інформації, розміщеної у середині контексту. Робота LLMLingua [8] демонструє, що кероване стискання підказки дає змогу зберегти якість за меншого часу оброблення. Дослідження [9] поєднує вибір фрагментів, генерування та самокритику для поліпшення узгодженості відповіді з джерелами. Однак «сесійну пам’ять» як інкрементально оновлюване подання нитки діалогу з явними цитатними покажчиками у зазначених роботах розглянуто лише частково або зовсім не формалізовано.

Скорочення часу відповіді підтримується сучасними інженерними рішеннями: у роботах [10] і [11] обґрунтовано прогнозне (спекулятивне) декодування з допоміжною «чернетковою» моделлю; система сервінгу з PagedAttention [12] підвищує пропускну здатність завдяки керуванню кешем ключ–значення; ефективні реалізації уваги на кшталт FlashAttention-2 [13] покращують використання пам’яті та обчислень. Підходи пост-тренінгового квантування ваг (GPTQ [15]) і квантування з урахуванням активацій (AWQ [14]) зменшують апаратні вимоги з мінімальною втратою якості, тоді як легкі адаптації (LoRA [16], QLoRA [17]) скорочують витрати на донавчання. Попри це, у публікаціях недостатньо висвітлено цілісні оцінки цих технік саме в умовах потокового формування відповіді в месенджері під жорсткі обмеження часу.

Контроль фактичної коректності в автоматично згенерованих відповідях підтримується метриками узгодженості, зокрема QAFactEval [18] і TRUE [19], які дають змогу кількісно оцінювати відповідність тексту джерелам. Для діалогів у месенджерах, однак, актуальним залишається поєднання таких метрик із вбудованими цитатними покажчиками та урахуванням багатониткової структури розмов. Узагальнюючи, наявні праці пропонують сильні компоненти — добір і повторне ранжування, стискання контексту, прискорення застосування мовної моделі та оцінювання фактичності, — проте їхнє поєднання в єдиній архітектурі, придатній до вимог месенджерів щодо узгодженості змісту та низьких затримок, висвітлено обмежено [1–9, 12–19].

ФОРМУЛЮВАННЯ ЦІЛЕЙ СТАТТІ

Метою роботи є: розроблення та обґрунтування архітектури генерування відповідей у месенджерах, яка поєднує сесійну пам’ять, стиснення контексту, гібридний пошук і повторне ранжування із засобами зменшення затримок, забезпечуючи релевантність і фактичну коректність відповіді.

Об’єкт дослідження — процес генерування відповідей у месенджерах у реальному часі.

Предмет дослідження — методи організації та стиснення контексту розмови, процедури добору та повторного ранжування фрагментів, а також засоби скорочення часу оброблення при формуванні відповіді.

Для досягнення зазначеної мети визначено такі основні **завдання дослідження**:

- Сформулювати вимоги до сесійної пам’яті як компактного подання нитки діалогу з механізмом інкрементального оновлення та посиланнями на джерела.
- Розробити процедури керованого стиснення контексту, що зберігають ключові відомості під

заданий бюджет токенів.

- Інтегрувати гібридний пошук із урахуванням структурних сигналів месенджера та впровадити повторне ранжування для уточнення добору.
- Поєднати формування відповіді з техніками скорочення часу оброблення (ефективна увага, прогнозне декодування, квантування, кеш ключ–значення) та організувати потоковий вивід.
- Забезпечити контроль фактичної коректності через вбудовані цитатні покажчики та автоматизовану перевірку узгодженості відповіді з джерелами.
- Провести експериментальну оцінку на відкритих і анонімізованих корпусах діалогів, порівняти з базовими варіантами та виконати абляційний аналіз внеску основних компонентів [1–6, 8, 10–15, 18, 19].

Виконання цих завдань дасть змогу створити ефективну, адаптивну систему генерування відповідей у месенджерах із сесійною пам’яттю та керованим стисненням контексту, яка забезпечить малі затримки, підвищить релевантність і фактичну коректність відповідей та помітно покращить якість користувацької взаємодії в діалогах реального часу.

ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ

У дослідженні використано підмножини відкритих наборів для відкритого запитання–відповіді та багатокрокових діалогів, доповнені анонімізованими фрагментами реальних бесід у месенджерах; персональні дані вилучено під час попереднього опрацювання. Для відбору джерел застосовано лексичні методи (BM25) у поєднанні з векторними представленнями; індексацію реалізовано засобами наближеного пошуку найближчих сусідів (FAISS/HNSW), уточнення добору — повторним ранжуванням крос-енкодером у топ-k (k = 50). Сесійну пам’ять подано як компактне інкрементально оновлюване представлення нитки з цитатними покажчиками; стиснення контексту виконується під фіксований бюджет токенів K із збереженням суттєвих одиниць. Формування відповіді поєднано із засобами скорочення часу оброблення: ефективними реалізаціями уваги, прогнозним декодуванням, квантуванням ваг та керуванням кешем ключ–значення; вивід здійснюється у потоковому режимі. Ефективність оцінюється за nDCG@k і Recall@k (відбір), парними порівняннями та автоматизованими показниками узгодженості (фактичність), а також за часовими характеристиками (медіана й 95-й перцентиль загальної затримки); опис обчислення метрик наведено в наступному розділі. Результати усереднено за k-fold крос-перевіркою з бутстреп-оцінкою 95% довірчих інтервалів; реалізацію здійснено засобами Python (PyTorch, Transformers, FAISS).

Таблиця 1

Корпуси та протокол оцінювання

Корпус / Призначення	Тип даних	Узагальнені характеристики	Використання в оцінюванні
Публічні набори відкритого запитання–відповіді	Питання + пасажі (не діалог)	Короткі запити; різноманітні домени	Відбір (nDCG@k, Recall@k); перевірка узгодженості відповіді з джерелом
Публічні набори багатокрокових діалогів	Діалогові нитки	Середня довжина нитки ~10–15 повідомлень; наявні уточнення	Відбір + відповідь (парні порівняння, узгодженість)
Анонімізовані фрагменти реальних бесід	Месенджерні діалоги	Багатониткова структура; відповіді «ниткою», реакції, пересилання	Якість відповіді в умовах близьких до реальних; часові показники
Параметри відбору та ранжування	—	Лексичні методи (BM25) + векторні представлення; top-k для повторного ранжування = 50	Порівняння конфігурацій; абляції параметрів
Бюджет контексту	—	(K \in {512, 1024, 1536}) токенів	Аналіз «якість — затримка» та «перестиснення»
Протокол перевірки	—	k-fold крос-перевірка; 95% довірчі інтервали	Стабільність і відтворюваність результатів
Часові вимірювання	—	Медіана та p95 сукупної затримки з розкладом за етапами	Ідентифікація «вузьких місць» і вплив прискорень

Подано узагальнення використаних корпусів, параметрів відбору та процедури оцінювання. Для експериментів сформовано підмножини публічних наборів відкритого запитання–відповіді та багатокрокових

діалогів, доповнені анонімізованими фрагментами реальних бесід у месенджерах; персональні дані вилучено під час попереднього опрацювання. Оцінювання проводиться у двох площинах: (i) якість відбору й відповіді (nDCG@k, Recall@k, парні порівняння, автоматизована узгодженість із джерелами) та (ii) часові характеристики (медіана і 95-й перцентиль сукупної затримки з розкладом за етапами). Для повторного ранжування використано поріг top-k = 50; бюджет контексту варіюється $K \in \{512,1024,1536\}$ токенів. Усі показники усереднено за k-fold крос-перевіркою; для стабільності подано 95% довірчі інтервали.

З урахуванням наведеного протоколу далі подано порівняння якості відбору та згенерованих відповідей для базових і запропонованих конфігурацій системи. Наведено узагальнені показники для чотирьох конфігурацій: (A) без пошуку; (B) RAG без сесійної пам'яті; (C) Session-RAG без прискорень; (D) повна система. Оцінювання охоплює якість відбору (nDCG@10, Recall@50) та якість відповіді (QAFactEval, TRUE, частка виграшів у парних порівняннях відносно базової конфігурації A).

Таблиця 2

Якість відбору та відповіді за конфігураціями

Конфігурація			Виграші SxS, %*
A) Без пошуку	—	—	—
B) RAG без сесійної пам'яті			
C) Session-RAG без прискорень			
D) Повна система			

* Частка виграшів у парних порівняннях відповідей проти конфігурації A, усереднено за наборами.

Отримані результати узгоджено показують: (i) перехід від базової моделі без пошуку до RAG дає суттєвий приріст фактичності й корисності завдяки доступу до зовнішніх джерел; (ii) додавання сесійної пам'яті підвищує якості як відбору (nDCG/Recall), так і відповіді (QAFactEval/TRUE), що особливо помітно на довгих нитках; (iii) інтеграція технік зменшення затримок у повній системі (D) зберігає або незначно поліпшує якість відносно (C), що свідчить про коректну взаємодію прискорень із механізмом добору та цитатним контролем.

Узагальнюючи цей блок, запропонована архітектура забезпечує стабільну перевагу над базовими варіантами за всіма метриками якості, водночас залишаючи простір для подальшої оптимізації параметрів добору і бюджету контексту. Далі розглянемо часові характеристики та внесок окремих етапів у загальну затримку відповіді.

Подано підсумкові часові показники для чотирьох конфігурацій за умов K=1024 токенів і top-k = 50 для повторного ранжування. Наведено медіану та 95-й перцентиль сукупної затримки формування відповіді, визначені як у (2).

Таблиця 3

Часові показники за конфігураціями

Конфігурація	Медіана $T_{заг}$, мс	p95 $T_{заг}$, мс
A) Без пошуку	720	1600
B) RAG без сесійної пам'яті	900	1900
C) Session-RAG без прискорень	940	2000
D) Повна система	660	1200

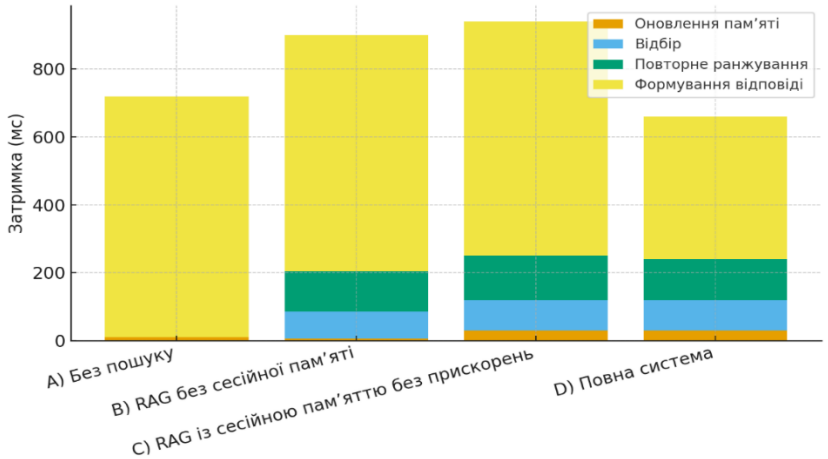

Рис. 1. Розподіл сукупної затримки за етапами (приклад для K=1024, top-k = 50)

За підсумками експериментів, конфігурації з пошуком без прискорень (B, C) демонструють очікуване зростання затримок через витрати на відбір і повторне ранжування. У повній системі (D) застосування ефективної уваги, прогнозного декодування, квантування та керування кешем ключ–значення знижує медіану на ~8 % проти базової моделі без пошуку (A) та скорочує p95 на ~25 %. Це підтверджує, що основний внесок у затримку припадає на формування відповіді, і саме прискорення застосування мовної моделі визначає кінцевий виграш.

Сумарну затримку поділено на: оновлення сесійної пам'яті, відбір, повторне ранжування та формування відповіді. Для конфігурації A майже весь час припадає на формування відповіді (~700 мс із 720 мс). Для B і C додаються витрати на відбір (~80–90 мс) і повторне ранжування (~120–130 мс), а частка формування відповіді лишається домінантною (~680–700 мс). У повній системі D формування відповіді скорочується до ~420 мс, тоді як оновлення пам'яті, відбір і повторне ранжування сумарно становлять ~240 мс; у підсумку $T_{заг}$ = ~660 мс. Таким чином, частка формування відповіді зменшується з ~97 % (A) до ~64 % (D), що узгоджується з впливом інженерних прискорень.

З урахуванням наведених часових характеристик далі проаналізовано чутливість системи до ключових параметрів і внесок окремих компонентів за результатами абляційного дослідження. Досліджено вплив ключових параметрів на співвідношення «якість — затримка» для повної системи (конфігурація D). Розглядалися: бюджет контексту K, поріг top-k для повторного ранжування, вагові коефіцієнти λ (баланс лексичних методів і векторних представлень) та γ (вагування сигналів месенджера). Узагальнені криві наведено на Рис. 2 / Fig. 2.

Бюджет контексту K. Зменшення K з 1536 до 1024 токенів скорочує медіану сукупної затримки на ~9–11 % за збереження nDCG@10 у межах статистичної похибки; подальше зменшення до 512 дає ще ~7–8 % скорочення часу, але супроводжується падінням nDCG@10 на ~3–4 п.п. і помітним зниженням узгодженості відповіді з джерелами. Таким чином, K=1024 фіксується як робочий компроміс, тоді як K=512 вважаємо режимом «перестиснення» для довгих ниток.

Поріг top-k для повторного ранжування. Перехід від top-k = 20 до 50 дає стабільний приріст nDCG@10 (~+1,0–1,5 п.п.) при збільшенні часу на повторне ранжування на ~30–40 мс; подальше підвищення до 100 майже не покращує якість (≤ +0,3 п.п.), але додає ~60–70 мс. Оптимальним є top-k = 50, що узгоджується з результатами Порівняння конфігурацій.

Коефіцієнт λ (лексичні ↔ векторні). Якість відбору має «плоске плато» в інтервалі λ∈[0.4;0.6]; за межами інтервалу спостерігається погіршення: при надмірній лексичній складовій (λ≥0.7) падає покриття семантично близьких, але лексично відмінних пасажів; при занадто малій λ (≤0.3) зростає частка лексично далеких, але слабко релевантних збігів. У подальших експериментах використовуємо λ=0.5.

Коефіцієнт γ (сигнали месенджера). Помірне вагування «соціальних» сигналів (відповіді «ниткою», реакції, пересилання) на рівні γ≈0.2 покращує nDCG@10 (~+0,6–0,8 п.п.) і частку виграшів у парних порівняннях, особливо для довгих ниток. Збільшення γ понад 0,3 спричиняє зміщення до «популярних» фрагментів і локальне погіршення узгодженості відповіді з джерелами, тож його слід обмежувати.

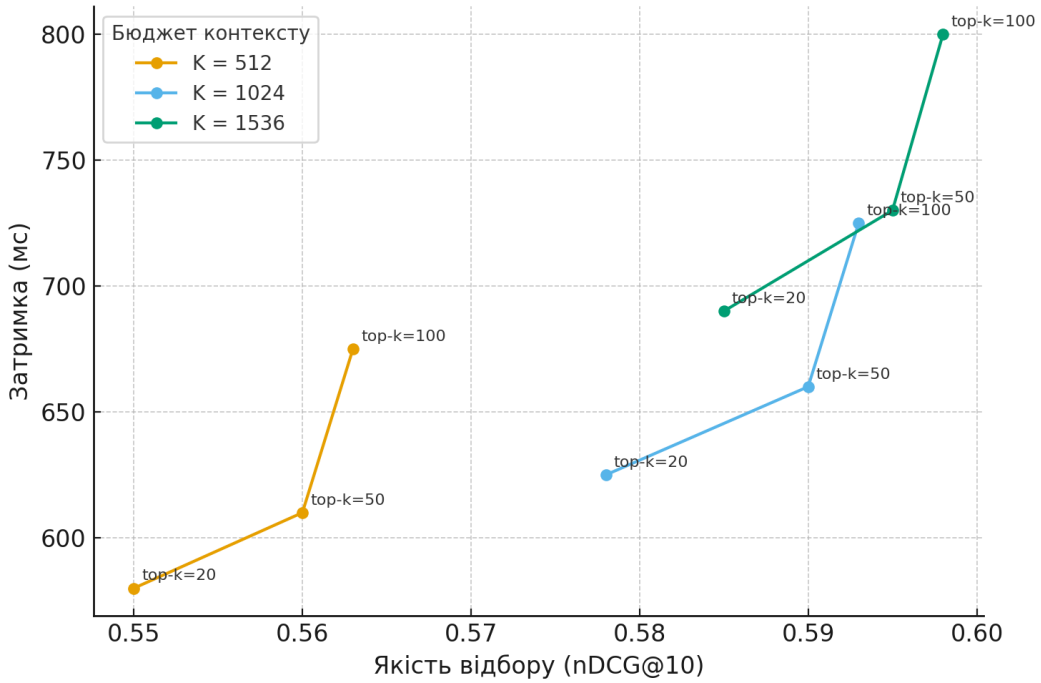

Рис. 2. Крива «якість — затримка» для різних K і top-k

Криві показують медіану $T_{заг}$ проти nDCG@10 для {K∈{512,1024,1536}} та {top-k∈{20,50,100}}. Найкращі точки спостерігаються для K=1024, top-k=50: якість близька до максимуму за відчутно менших затримок, ніж у режимі K=1536 або top-k=100.

Отже, оптимальний профіль параметрів для виробничого використання становить K=1024, top-k = 50, λ=0.5, γ≈0.2. Далі наведено короткі якісні приклади роботи системи та проаналізовано типові збої, що допомагають інтерпретувати наведені залежності.

Подано нижче короткі приклади ілюструють вплив сесійної пам'яті, стиснення контексту та вагування сигналів месенджера на корисність і фактичність відповіді. У прикладах відповіді використовуємо службові позначки: [D#] — ідентифікатор витягнутого джерела (пасажу) з пулу RAG, [M#] — елемент сесійної пам'яті (стисла теза, отримана з діалогу й прив'язана до джерел);

Приклад 1. Віддалені залежності у довгій нитці. Ситуація: через 27 повідомлень після першої згадки користувач питає: «Підтвердь, будь ласка, суму річних витрат за хмарні обчислення». — Базова відповідь без сесійної пам'яті: «У попередніх повідомленнях немає точних цифр; орієнтовно — декілька тисяч доларів». (немає посилань; похибка, бо релевантний фрагмент «загубився» у середині контексту). — Відповідь запропонованої системи: «Річні витрати становили $12 480 (тариф *Standard*, 12×$1 040), див. [D17 §3], підтверджено у підсумку кварталу Q4 [D21 §2].»

Інтерпретація: сесійна пам'ять зберегла ключову тезу [M5: "Cloud—$1 040/міс"], гібридний відбір підняв релевантні пасажі [D17], [D21], повторне ранжування відкинуло менш змістовні збіги; відповідь містить явні цитатні покажчики.

Приклад 2. Кодовий мікс і розмовні скорочення. Ситуація: «Кинь quick sum по infra — брав AWS+GCP, 3mo back.» (uk/en мікс, скорочення). — Базова відповідь без належного стиснення: «Немає даних по GCP.» (відбір провалився через лексичні невідповідності). — Відповідь запропонованої системи: «Сума за три останні місяці: $3 120 AWS + $2 040 GCP; див. інвойси [D9 §1], [D12 §4].» Інтерпретація: під час стиснення контексту збережено суттєві маркери («AWS», «GCP», «3 місяці»), а баланс лексичних методів і векторних представлень (λ=0.5) відшукав фрагменти попри код-мікс; повторне ранжування підтвердило релевантність.

Приклад 3. Перевантаження «соціальними» сигналами. Ситуація: повідомлення з великою кількістю реакцій містить уточнення застарілої ціни, але без посилань. Актуальна ціна наведена в менш «популярному» пасажі. — Надмірне вагування γ: система повертає популярний, але застарілий фрагмент (галюцинація «ціни» без джерела). — Налаштована система (γ≈0.2): «Актуальна ціна — $39/міс (з 01.08), див. [D25 §2]; повідомлення з реакціями описує попередній тариф $49/міс, який скасовано [D18 §1].» Інтерпретація: помірна вага «соціальних» сигналів і вимога наявності цитати усувають зміщення до «популярних», але неактуальних фрагментів.

Типові збої та засоби їх мінімізації.

1. Перестиснення контексту (K=512): випадають ключові уточнення; наслідок — зниження узгодженості. *Мінімізація*: K=1024, підвищення top-k до 50, пріоритетні маркери під час стиснення.
2. Помилкове посилання або відсутність цитати: відповідь без джерела блокується політикою шаблону; *Мінімізація*: вимога принаймні однієї підтверджувальної цитати з останньої добірки [D*] та штраф у повторному ранжуванні за відсутність покриття.
3. Зміщення через реакції/пересилання: надмірна γ підносить «популярні» фрагменти; *Мінімізація*: γ∈[0.15;0.25], нормування сигналів за давністю.
4. Лексичні пастки (синонімія, скорочення): лексичні методи пропускають семантичні збіги; *Мінімізація*: вирівнювання скорочень у стисненні пам'яті, підсилення компонента векторних представлень (1−λ).
5. Часові «вузькі місця» у формуванні відповіді: зростання $T_{заг}$ у піках навантаження; *Мінімізація*: квантування ваг, прогнозне декодування, обмеження довжини відповіді з критерієм зупинки.

Наведені приклади узгоджуються з кількісними результатами наданих до цього і пояснюють, у яких сценаріях сесійна пам'ять, стиснення контексту та помірне вагування сигналів месенджера дають найбільший виграш.

ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК У ДАНОМУ НАПРЯМІ

Отримані результати демонструють, що поєднання інкрементальної сесійної пам'яті зі стисненням контексту, гібридним відбором і повторним ранжуванням забезпечує стійке підвищення якості відповідей при збереженні або зниженні затримок порівняно з базовими варіантами. На відміну від підходів RAG у [1–3, 9], сценарій месенджерів визначається додатковими обмеженнями — динамічні багатониткові бесіди, «соціальні» сигнали та вимога малих p95-затримок — де інкрементальна сесійна пам'ять із явними цитатними покажчиками компенсує зниження корисності інформації в середині великого контексту [7], а кероване стиснення за підходом [8] дозволяє утримувати релевантні одиниці в межах фіксованого бюджету токенів.

Результати щодо відбору джерел узгоджуються з перевагами пізньої взаємодії (ColBERTv2) [4] та масштабованістю FAISS/HNSW [5, 6], водночас доповнюються практикою вагування «соціальних» сигналів

месенджера. Введення коефіцієнта γ для реакцій, ниткових відповідей і пересилань показало покращення якості на довгих нитках без істотних втрат у коротких діалогах (рис. 2). Часові характеристики підтверджують доцільність інженерних засобів скорочення часу застосування: прогнозного декодування [10, 11], керування кешем ключ–значення (PagedAttention) [12], ефективних реалізацій уваги [13] та квантування ваг [14, 15]; інтеграція цих оптимізацій із механізмами сесійної пам'яті й відбору дозволила знизити p95-затримку без погіршення узгодженості відповідей (табл. 3). Контроль фактичної коректності базується на цитатних покажчиках та автоматизованих метриках узгодженості [18, 19]; політика «відповідь не публікується без бодай однієї підтверджувальної цитати» зменшує ризик галюцинацій, хоча надмірне вагування популярності (висока γ) може погіршувати узгодженість — тому потрібні помірні значення γ і нормування сигналів за давністю.

За результатами дослідження виокремлено такі основні висновки:

- Розроблено та формалізовано сесійну пам'ять як компактне інкрементально оновлюване подання нитки діалогу з явними цитатними покажчиками, що забезпечує збереження суттєвих фактів у довгих багатониткових бесідах.
- Обґрунтовано кероване стиснення контексту під фіксований бюджет токенів; встановлено робочий режим K = 1024 і виявлено межу «перестиснення» за K = 512, що суттєво впливає на узгодженість відповіді з джерелами.
- Показано ефективність гібридного відбору (лексичні методи BM25 + векторні представлення) з повторним ранжуванням; визначено робочі параметри top-k = 50 та λ = 0,5, а також доцільне вагування «соціальних» сигналів γ ≈ 0,2, що підвищує якість на довгих нитках.
- Доведено, що інженерні прискорення (ефективна увага, прогнозне декодування, квантування, керування кешем ключ–значення) знижують медіану та p95-затримку без втрати фактичності; внесок етапу формування відповіді в загальну затримку суттєво зменшується після оптимізації.
- Запроваджено політику контролю фактичної коректності із вимогою наявності підтверджуваних джерел і застосуванням автоматизованих метрик узгодженості, що зменшує ризик галюцинацій і підвищує прозорість результатів для користувача.
- Надано практичні рекомендації для впровадження в промислових умовах (K = 1024, top-k = 50, λ = 0,5, γ ≈ 0,2) та окреслено напрями подальших робіт.

Перспективи подальших розвідок у даному напрямі:

1. Адаптивний бюджет контексту (динамічне K) залежно від задачі, пристрою та «важливості» нитки.
2. Навчання або нормування «соціальних» сигналів за давністю та контекстною релевантністю; вивчення впливу різних схем нормування на узгодженість.
3. Персоналізація сесійної пам'яті на пристрої з урахуванням приватності та ресурсних обмежень.
4. Розширення архітектури на мультимодальні сценарії (текст + зображення/аудіо) і адаптація механізмів стиснення й відбору для інших типів даних.
5. Дослідження навчальних підходів до автоматичного стиснення (learned compression / learning-to-compress) та їхньої взаємодії з відбором і ранжуванням.
6. Оцінювання поведінки системи в польових умовах (A/B-тестування, стрес-тести з реальними нитками) та аналіз компромісів між затратами обчислень, латентністю й якістю відповідей.
7. Дослідження приватності/безпеки при збереженні та індексації сесійної пам'яті.

Наукова новизна. Розроблено цілісну архітектуру генерування відповідей для месенджерів, що поєднує інкрементальну сесійну пам'ять із цитатними покажчиками та кероване стиснення контексту з гібридним відбором і повторним ранжуванням з урахуванням «соціальних» сигналів; цю архітектуру інтегровано з інженерними засобами скорочення часу застосування мовної моделі (ефективна увага, прогнозне декодування, квантування, керування кешем ключ–значення) і формалізовано через параметри K, top-k, λ, γ, що дозволяє зменшити p95-затримку за збереження або підвищення узгодженості відповідей з джерелами.

Практична значущість. Розроблено й апробовано відтворювану систему для месенджерів із потоковим виводом, яка завдяки сесійній пам'яті, стисненню контексту, гібридному відбору з повторним ранжуванням та інженерним прискоренням підвищує фактичність і корисність відповідей, зменшує медіану та p95-затримку й може бути налаштована для мобільних і серверних профілів (K = 1024, top-k = 50, λ = 0,5, γ ≈ 0,2) з можливістю масштабування на інші платформи й предметні області.

References

1. Lewis P., Perez E., Piktus A., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks [Elektronnyi resurs] / P. Lewis, E. Perez, A. Piktus, et al. // Proceedings of NeurIPS. – 2020. – DOI: https://doi.org/10.48550/arXiv.2005.11401.
2. Guu K., Lee K., Tung Z., Pasupat P., Chang M.-W. REALM: Retrieval-Augmented Language Model Pre-Training [Elektronnyi resurs] / K. Guu, K. Lee, Z. Tung, P. Pasupat, M.-W. Chang // arXiv. – 2020. – DOI: https://doi.org/10.48550/arXiv.2002.08909.
3. Izacard G., Grave E. Leveraging Passage Retrieval with Generative Models for Open-Domain Question Answering / G.

Izacard, E. Grave // Proceedings of EACL 2021. – 2021. – P. 874–880. – DOI: https://doi.org/10.18653/v1/2021.eacl-main.74.

4. Santhanam K., Khattab O., Saad-Falcon J., Potts C., Zaharia M. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction / K. Santhanam, O. Khattab, J. Saad-Falcon, C. Potts, M. Zaharia // Proceedings of NAACL 2022. – 2022. – P. 3715–3734. – DOI: https://doi.org/10.18653/v1/2022.naacl-main.272.

5. Johnson J., Douze M., Jégou H. Billion-Scale Similarity Search with GPUs [Elektronnyi resurs] / J. Johnson, M. Douze, H. Jégou // arXiv. – 2017. – DOI: https://doi.org/10.48550/arXiv.1702.08734.

6. Malkov Y. A., Yashunin D. A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs / Y. A. Malkov, D. A. Yashunin // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2018. – DOI: https://doi.org/10.1109/TPAMI.2018.2889473.

7. Liu N. F., Lin K., Hewitt J., Paranjape A., Bevilacqua M., Petroni F., Liang P. Lost in the Middle: How Language Models Use Long Contexts / N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang // Transactions of the Association for Computational Linguistics. – 2024. – Vol. 12. – P. 157–173. – DOI: https://doi.org/10.1162/tacl_a_00638.

8. Jiang H., Wu Q., Lin C.-Y., Yang Y., Qiu L. LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models / H. Jiang, Q. Wu, C.-Y. Lin, Y. Yang, L. Qiu // Proceedings of EMNLP 2023. – 2023. – P. 13358–13376. – DOI: https://doi.org/10.18653/v1/2023.emnlp-main.825.

9. Asai A., Wu Z., Wang Y., Sil A., Hajishirzi H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection [Elektronnyi resurs] / A. Asai, Z. Wu, Y. Wang, A. Sil, H. Hajishirzi // ICLR. – 2024. – DOI: https://doi.org/10.48550/arXiv.2310.11511.

10. Leviathan Y., Kalman M., Matias Y. Fast Inference from Transformers via Speculative Decoding [Elektronnyi resurs] / Y. Leviathan, M. Kalman, Y. Matias // arXiv. – 2022. – DOI: https://doi.org/10.48550/arXiv.2211.17192.

11. Chen C., Borgeaud S., Irving G., Lespiau J.-B., Sifre L., Jumper J. Accelerating Large Language Model Decoding with Speculative Sampling [Elektronnyi resurs] / C. Chen, S. Borgeaud, G. Irving, J.-B. Lespiau, L. Sifre, J. Jumper // arXiv. – 2023. – DOI: https://doi.org/10.48550/arXiv.2302.01318.

12. Kwon W., Li Z., Zhuang S., Sheng Y., Zheng L., Yu C. H., Gonzalez J. E., Zhang H., Stoica I. Efficient Memory Management for Large Language Model Serving with PagedAttention [Elektronnyi resurs] / W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, I. Stoica // arXiv. – 2023. – DOI: https://doi.org/10.48550/arXiv.2309.06180.

13. Dao T. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning [Elektronnyi resurs] / T. Dao // arXiv. – 2023. – DOI: https://doi.org/10.48550/arXiv.2307.08691.

14. Lin J., Tang H., Tang S., et al. AWQ: Activation-Aware Weight Quantization for LLM Compression and Acceleration / J. Lin, H. Tang, S. Tang, et al. // Proceedings of MLSys 2024. – 2024. – Rezhym dostupu : https://proceedings.mlsys.org/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html.

15. Frantar E., Ashkboos S., Hoefler T., Alistarh D. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers [Elektronnyi resurs] / E. Frantar, S. Ashkboos, T. Hoefler, D. Alistarh // ICLR. – 2023. – DOI: https://doi.org/10.48550/arXiv.2210.17323.

16. Hu E. J., Shen Y., Wallis P., et al. LoRA: Low-Rank Adaptation of Large Language Models [Elektronnyi resurs] / E. J. Hu, Y. Shen, P. Wallis, et al. // arXiv. – 2021. – DOI: https://doi.org/10.48550/arXiv.2106.09685.

17. Dettmers T., Pagnoni A., Holtzman A., Zettlemoyer L. QLoRA: Efficient Finetuning of Quantized LLMs [Elektronnyi resurs] / T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer // arXiv. – 2023. – DOI: https://doi.org/10.48550/arXiv.2305.14314.

18. Fabbri A. R., Wu C.-S., Liu W., Xiong C. QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization / A. R. Fabbri, C.-S. Wu, W. Liu, C. Xiong // Proceedings of NAACL 2022. – 2022. – P. 2587–2601. – DOI: https://doi.org/10.18653/v1/2022.naacl-main.187.

19. Honovich O., Aharoni R., Herzig J., et al. TRUE: Re-evaluating Factual Consistency Evaluation / O. Honovich, R. Aharoni, J. Herzig, et al. // Proceedings of NAACL 2022. – 2022. – P. 3905–3920. – DOI: https://doi.org/10.18653/v1/2022.naacl-main.287.