foxops TECH

ИССЛЕДОВАНИЕ · ПРЕПРИНТ

Детерминированная долговременная память для LLM-агентов

Гибридный поиск, документный граф связей и масштабируемость записи в изолированном контуре

Следнев А. А. · Slednev A.

ООО «ФОКСОПС» / LLC FOXOPS Tech, Moscow, Russia
ORCID https://orcid.org/0009-0002-7714-4769 · aslednev@foxops.tech

Preprint · Version 4.0 · Published 2026-07-05 · DOI: 10.5281/zenodo.21210732

Детерминированная долговременная память для LLM-агентов: гибридный поиск, документный граф связей и масштабируемость записи

Следнев А. А. · ООО «ФОКСОПС» · aslednev@foxops.tech

Аннотация. Системы долговременной памяти для агентов на основе больших языковых моделей (LLM), как правило, используют обращения к языковой модели на пути записи: извлечение фактов и сущностей, разрешение кореференции, согласование противоречий. В работе исследуется альтернативная точка проектного пространства — память, путь записи которой не содержит генеративных обращений к LLM (извлечения фактов и сущностей); плотный поиск использует нейросетевой эмбеддер. Источником истины служат версионируемые Markdown-документы; над ними детерминированными алгоритмами строится производный индекс: полнотекстовый поиск с учётом русской морфологии, плотные векторные представления и документный граф связей, объединяющий явные вики-ссылки с автоматически порождаемыми «несвязанными упоминаниями», устойчивыми к словоизменению. На двух корпусах (29 документов / 100 вопросов; 68 документов / 50 вопросов) относительно вручную выверенного референса гибридный протокол достигает hit@5 88–89% и полноты multi-hop 93–100%; графовая экспансия повышает полноту multi-hop с 27–33% до 73–80%. Базовая система на основе LLM-экстракции (Graphiti; Qwen2.5-7B-Instruct, Q4_K_M) достигает 42% (67% на уровне фактов) при времени индексации примерно в 700 раз большем. На открытом корпусе из 3600 статей показано, что исходная реализация содержит на пути записи операции, линейные по числу неразрешённых (ghost) рёбер, что даёт выраженный сверхлинейный рост суммарной загрузки; деградация локализована до трёх запросов сложности O(N) и устранена индексами и переносом нормализации в SQL, после чего запись масштабируется практически линейно (ускорение в 3,7 раза при 3600 документах при сохранении семантики связей).

Ключевые слова: долговременная память агентов; retrieval-augmented generation; гибридный поиск; документный граф знаний; несвязанные упоминания; детерминированная индексация; масштабируемость; PostgreSQL; pgvector; изолированный контур.

1 Введение

Агент, перечитывающий первоисточники при каждом запросе, оплачивает одно и то же знание многократно — токенами, латентностью и объёмом контекстного окна. Стандартный подход, генерация с дополнением поиском (retrieval-augmented generation, RAG) [1], оставляет знание в сырых документах и переоткрывает его на каждом запросе. Идея однократной «компиляции» корпуса в постоянную базу знаний восходит к концепции Memex В. Буша [2] и недавно вернулась в практику в виде LLM-поддерживаемых персональных вики [3]. Современные системы агентной памяти — MemGPT [4], Mem0 [5], Zep/Graphiti [6] — реализуют этот процесс обращениями к LLM на пути записи.

У такого дизайна два следствия, существенных для изолированных (air-gapped) контуров. Экономическое: стоимость индексации пропорциональна размеру корпуса в обращениях к модели, а доступны лишь локальные модели ограниченного размера. Эпистемологическое: хранилище, наполняемое генеративной моделью, может содержать утверждения без соответствия во входных данных, чей провенанс не прослеживается до источника.

Цель работы — исследовать противоположную точку проектного пространства как проектное ограничение: *путь записи не содержит генеративных компонентов*. Вся генеративная работа выносится к агенту-читателю. Вклад: (1) архитектура памяти с версионируемым человекочитаемым слоем истины и детерминированным производным индексом; (2) документный граф с автоматическими упоминаниями, устойчивыми к морфологии; (3) эмпирическое сравнение режимов поиска и базовой системы с LLM-экстракцией; (4) исследование масштабируемости на открытом корпусе из 3600 документов — основное дополнение версии.

2 Материалы и методы

2.1 Архитектура системы

Система состоит из слоя истины и производных слоёв; инвариант — производные слои полностью перестраиваемы из слоя истины детерминированной переиндексацией. Каждое пространство знаний — git-репозиторий Markdown-документов. Запись атомарна: git-фиксация, затем одна транзакция реляционного индекса (PostgreSQL 16). Уникальность — в границе генеративности (рис. 1): всё внутри пунктирного контура детерминировано и не содержит генеративной LLM; языковая модель работает только у агента-читателя. Отсюда три отличия от систем на LLM-экстракции: слой истины хранит документы побайтно; граф строится разрешением авторских вики-ссылок и морфологическим сопоставлением имён (рис. 2), а не извлечением сущностей; стоимость индексации не зависит от числа обращений к модели.

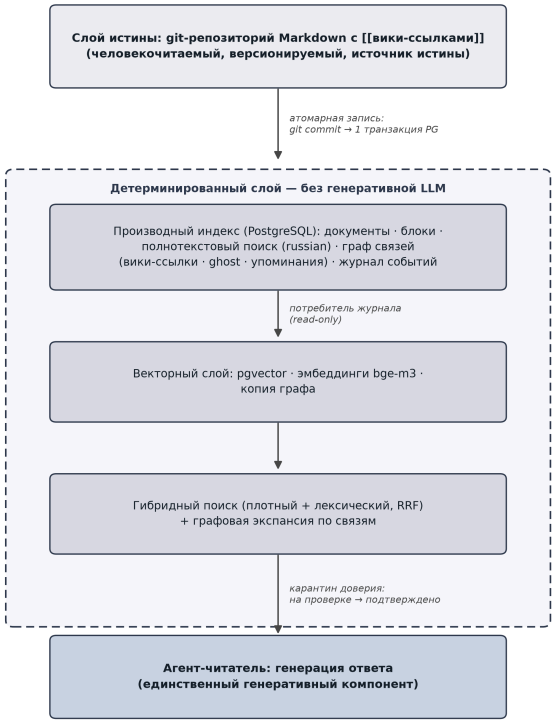

Рис. 1. Архитектура: детерминированный слой без генеративной LLM; генерация — только у агента.

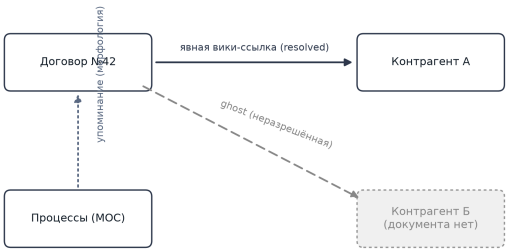

Рис. 2. Документный граф: явная вики-ссылка, ghost и упоминание (морфология).

Каждая запись порождает событие с детерминированным идентификатором в журнал (transactional outbox); потребитель журнала строит векторный слой (эмбеддинги bge-m3 [13], pgvector) и копию графа. Полный поток данных при записи и построение связей — на рис. 3.

Рис. 3. Поток данных при записи и построение связей; внизу — разрешение ghost-ребра во времени (HEAL).

Карантин доверия и провенанс. Новые и изменённые документы получают статус «на проверке»; в выдачу генерирующих потребителей идут только подтверждённые человеком. Генеративные утверждения без провенанса исключены конструктивно, однако семантическая истинность каждого автоматического ребра не гарантирована (омонимия, широкие совпадения); точность автосвязывания — предмет отдельной оценки.

2.2 Гибридный поиск и метрики

Запрос исполняется плотным (косинус эмбеддингов) и лексическим (полнотекстовый) пулами со слиянием Reciprocal Rank Fusion [12], константа сглаживания $k_0 = 60$:

$$\mathrm{RRF}(d) = \sum_{r \in R} \frac{1}{k_0 + \mathrm{rank}_r(d)}$$

Метрики: hit@k — доля вопросов, для которых документ-ответ d^* в топ-k; MRR — средний обратный ранг; полнота multi-hop — доля найденных документов ответа:

$$\mathrm{hit@}k = \frac{1}{|Q|} \sum_{q \in Q} \mathbb{1}\left[\mathrm{rank}_q(d_q^*) \le k\right]$$

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\mathrm{rank}_q(d_q^*)}$$

$$\mathrm{Recall}_{\mathrm{MH}}(q) = \frac{|D_{\mathrm{retr}}(q) \cap D_q^*|}{|D_q^*|}$$

Метрики hit@k и MRR определены для вопросов с документом-ответом d^*; негативные вопросы оцениваются отдельно — долей ложноположительной выдачи.

2.3 Корпуса и режимы

T1 — дистиллированная база знаний: 29 заметок по кодовой базе (65 рёбер). T2 — корпоративный корпус:

Режим	Термины	Парафразы	Multi-hop	Полнота MH
T1 · наивный FTS	7%	0%	0%	0%
T1 · итеративный FTS	87%	43%	83%	30%
T1 · семантика (сиды)	93%	60%	90%	33%
T1 · семантика + граф	93%	60%	90%	73%
T1 · объединение	**100%**	**67%**	**97%**	**93%**
T2 · итеративный FTS	73%	73%	87%	33%
T2 · семантика (сиды)	67%	80%	93%	27%
T2 · семантика + граф	67%	80%	93%	80%
T2 · объединение	**87%**	**80%**	**100%**	**100%**
T2 · Graphiti (док)	60%	20%	47%	7%
T2 · Graphiti (факт)	80%	47%	73%	—

Таблица 1. hit@5 по категориям вопросов и полнота multi-hop (MH). Референс = 100%.

68 документов (584 ребра, 8,6 ребра/документ; хранилище Obsidian). Вопросы четырёх типов: точные термины, парафразы, multi-hop, негативные; эталоны выверены чтением первоисточников, референс = 100%. Режимы: наивный и итеративный полнотекстовый; семантические сиды; графовая экспансия; объединение — целевой протокол агента.

2.4 Конфигурация и параметры

Компонент	Спецификация
Эмбеддер	bge-m3 [13]: 567M, F16, вектор 1024; Ollama 0.23.4
Экстрактор baseline	Qwen2.5-7B-Instruct, GGUF Q4_K_M; окно/контекст запуска 32768/16384
Инференс	temperature = 1, max_tokens = 16384 (defaults graphiti-core 0.29.2)
Baseline	Graphiti 0.29.2 [6] «из коробки»; FalkorDB
Наша система	PostgreSQL 16 + pgvector (HNSW, косинус); FTS russian; без генеративной LLM
Аппаратура	Apple M3 Max, Metal; модели локально через Ollama

2.5 Методика масштабирования

Открытый корпус — 3600 статей русской Википедии (CC BY-SA); вики-ссылки сохранены (⩽15/документ). Ступени 250/500/1000/2000/3600 (батчи по 50). Совокупная стоимость загрузки: G_i — накопленное число ghost-рёбер на момент записи i-го документа; до оптимизации $c_{\mathrm{pre}(i)}$ линейна по G_i, после — логарифмична:

$$T_{\mathrm{ingest}(N)} = \sum_{i=1}^{N} c(i), \quad c_{\mathrm{pre}(i)} = \Theta(G_i), \quad c_{\mathrm{opt}(i)} = \Theta(\log G_i)$$

Поскольку число ghost-рёбер растёт линейно с числом документов, сумма линейных стоимостей c_{pre} содержит квадратичный член, тогда как перевод горячих запросов в логарифмический доступ индексами (c_{opt}) делает загрузку практически линейной.

3 Результаты и обсуждение

3.1 Качество поиска

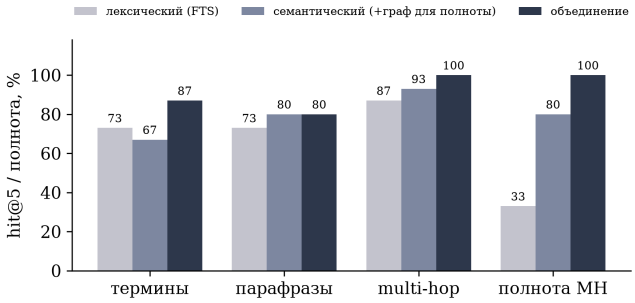

Рис. 4. Корпус T2: hit@5 и полнота multi-hop.

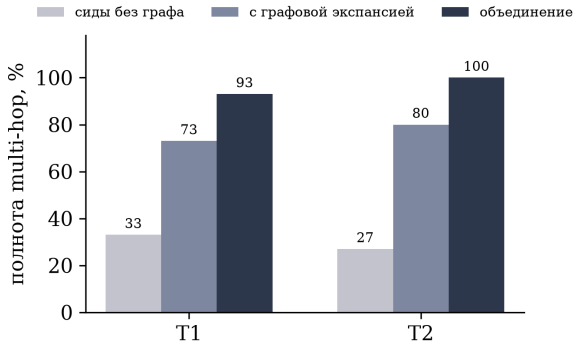

Рис. 5. Полнота multi-hop до и после графовой экспансии.

Три наблюдения. *Взаимодополняемость*: по MRR на T2 лексический превосходит семантический на терминах (0,66 против 0,53), семантический — на парафразах (0,57 против 0,37) и multi-hop (0,74 против 0,49); объединение не уступает лучшему одиночному режиму ни в одной категории. *Вклад графа* сосредоточен в полноте: экспансия повышает полноту multi-hop с 27–33% до 73–80% — второй документ ответа приходит по ребру графа. *Влияние хабов*: из 11 промахов семантики на T2 в 6 верх выдачи занят MOC-страницами.

Негативные вопросы: доля ложноположительных (FPR) 3/3 = 1,0 у семантики и 2/3 ≈ 0,67 у лексики (specificity ≈ 0 и ≈ 0,33). Выборка мала; отказ от ответа — обязанность генеративного слоя.

3.2 Стоимость индексации

Показатель	Наша	Graphiti
Индексация 68 док.	≈10 с	2 ч 04 мин
Обращений к ген. LLM	0	≈500
Повторная индексация	идентична	недетерм. (t=1)

При экстракции наблюдались фабрикации: искажённые имена сущностей и факты на смеси языков — строки, отсутствующие во входе. Эффект наблюдался в данном прогоне при бюджете локальной 7B-модели с Q4_K_M; с более сильным экстрактором частота снизится.

3.3 Масштабируемость записи

Наиболее значимый результат — на корпусе из 3600 документов. Исходная реализация: медианная одиночная запись возросла со 166 до 846 мс (в 5,1 раза), батч из 50 — с 4,3 до 34 с. Рост согласуется с O(N)-операциями по числу ghost-рёбер; чистый квадратичный закон не постулируется — рост сверхлинеен, но ниже квадратичной оценки. Чтение не деградировало: FTS 20–58 мс, семантика 173–256 мс, hit@5 семантики и TF-IDF = 100% на всех ступенях; pgvector корректен (HNSW, 5–19 мс против 188 мс перебором).

Диагностика планами (EXPLAIN ANALYZE) локализовала деградацию до трёх O(N)-запросов, два из которых не использовали индексы (нормализация имени в приложении). Устранение — миграция схемы (генерируемый столбец + частичные индексы, GIN по псевдонимам, индексы журнала) и переписывание OR на UNION ALL.

Запрос	До	После
HEAL (ghost-рёбра)	10,6 мс + 49 481 стр.	0,11 мс
Резолв на отсутств. док.	2,6 мс (скан)	0,16 мс
Согласование спикеров	2,8 мс (скан)	0,04 мс
Квота пространства	1,4 мс (скан)	0,9 мс

Таблица 2. Латентность горячих запросов записи (EXPLAIN ANALYZE, 49,5 тыс. рёбер).

После оптимизации повторный прогон показал устранение сверхлинейной деградации: время записи документа возросло со 107 до 147 мс (в 1,4 раза) против 87–550 мс (в 6,3 раза); при 3600 документах — ускорение в 3,7 раза по загрузке финальной ступени. Число ghost-рёбер совпало до единицы — семантика связей сохранена.

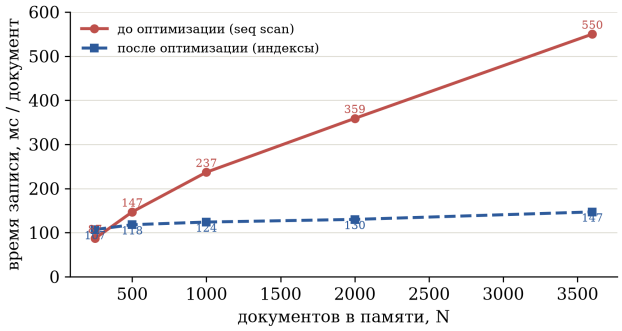

Рис. 6. Время записи документа: до оптимизации ×6,3, после — ×1,4.

Остаточный линейный рост (×1,4) — словарь имён для упоминаний (следующая оптимизация: кэш в процессе). Индексы чуть поднимают floor записи; точка окупаемости — около N = 400.

3.4 Угрозы валидности

Одноавторность: вопросы и эталоны — один исследователь; смещение может влиять и на относительный порядок режимов, проверка на независимой разметке — впереди. *Уровень метрики*: измерено качество поиска, не ответов. *Baseline*: Graphiti «из коробки», один прогон при temperature = 1, дисперсия не оценена; вывод — «в данной локальной конфигурации Graphiti уступила», не «уступает». Строгое сравнение требует 3–5 прогонов.

4 Заключение

Показано, что долговременная память для LLM-агентов может строиться без генеративных обращений к языковой модели на пути записи: детерминированная комбинация лексического поиска с морфологией, плотных представлений и документного графа достигает hit@5 88–89% и полноты multi-hop 93–100%, причём граф даёт основной прирост полноты (с 27–33% до 73–80%). Наивная реализация записи деградирует сверхлинейно; деградация локализуется до конкретных O(N)-запросов и устраняется стандартными средствами СУБД до линейного роста — без изменения семантики и без обращений к генеративной LLM. Границы: корпуса с авторской структурой, метрики уровня поиска, одноавторный датасет. Перспективы: битемпоральный слой; повторные прогоны baseline с оценкой дисперсии; прямая оценка точности автосвязывания; проверка на корпусах от десятков до сотен тысяч документов.

Литература

[1] Lewis P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks — NeurIPS 2020. arXiv:2005.11401.
[2] Bush V. As We May Think — The Atlantic Monthly. 1945. 176(1).
[3] Karpathy A. LLM Wiki. 2026.
[4] Packer C. et al. MemGPT: Towards LLMs as Operating Systems. 2023. arXiv:2310.08560.
[5] Chhikara P. et al. Mem0: Production-Ready AI Agents with Scalable Long-Term Memory. 2025. arXiv:2504.19413.
[6] Rasmussen P. et al. Zep: A Temporal Knowledge Graph Architecture for Agent Memory. 2025. arXiv:2501.13956.
[7] Edge D. et al. From Local to Global: A Graph RAG Approach. 2024. arXiv:2404.16130.
[8] Jiménez Gutiérrez B. et al. HippoRAG — NeurIPS 2024. arXiv:2405.14831.

[9] Robertson S., Zaragoza H. The Probabilistic Relevance Framework: BM25 and Beyond. 2009.
[10] Karpukhin V. et al. Dense Passage Retrieval — EMNLP 2020. arXiv:2004.04906.
[11] Thakur N. et al. BEIR — NeurIPS 2021 Datasets and Benchmarks. arXiv:2104.08663.
[12] Cormack G. V. et al. Reciprocal Rank Fusion — SIGIR 2009.
[13] Chen J. et al. M3-Embedding — Findings of ACL 2024. arXiv:2402.03216.
[14] Yang Z. et al. HotpotQA — EMNLP 2018. arXiv:1809.09600.
[15] Ho X. et al. 2WikiMultiHopQA — COLING 2020. arXiv:2011.01060.
[16] Trivedi H. et al. MuSiQue — TACL 2022. arXiv:2108.00573.
[17] Maharana A. et al. Evaluating Very Long-Term Conversational Memory — ACL 2024. arXiv:2402.17753.
[18] Wu D. et al. LongMemEval — ICLR 2025. arXiv:2410.10813.

foxops TECH

ИНФОРМАЦИЯ ОБ АВТОРЕ

Следнев Александр

ООО «ФОКСОПС» / LLC FOXOPS Tech — Moscow, Russia
организация-разработчик исследуемой системы · ИНН 7720961650 · ОГРН 1267700019014

ORCID https://orcid.org/0009-0002-7714-4769
E-MAIL aslednev@foxops.tech
WEB foxops.tech

RECOMMENDED CITATION

Slednev A. Deterministic Long-Term Memory for LLM Agents: Hybrid Retrieval, Document Graph, and Write-Path Scalability in Air-Gapped Environments. Preprint. Version 4.0. July 2026. DOI: https://doi.org/10.5281/zenodo.21210732.

Конфликт интересов. Автор — сотрудник ООО «ФОКСОПС», организации-разработчика исследуемой системы; сравнительная оценка проводилась на открытых и внутренних корпусах с вручную выверенным референсом.

Доступность данных. Открытый корпус масштабирования (статьи русской Википедии, CC BY-SA) и корпус T1 воспроизводимы. Артефакты воспроизводимости — генератор T1, вопросы и разметка, скрипты метрик, конфигурации PostgreSQL/Ollama/Graphiti, планы EXPLAIN и сырые тайминги — доступны от автора по запросу. Корпоративный корпус T2 конфиденциален и не публикуется.

foxops TECH

Обсудить пилот памяти — foxops.tech/contact