

ИНСТРУМЕНТ СОКРАЩЕНИЯ ОПЕРАЦИОННЫХ РАСХОДОВ WIKIAI

Разработана платформа WikiAi, интегрирующая RAG-технологии с графовыми базами знаний на основе микросервисной архитектуры. Предложена схема гибридного поиска с rank fusion. Проведена экспериментальная верификация на пилотных компаниях.

Введение

Сотрудники предприятий расходуют 25–30% рабочего времени на поиск корпоративной информации [1–3]. Традиционные системы управления знаниями не обеспечивают семантического понимания контекста. Парадигма RAG снижает галлюцинации языковых моделей за счёт интеграции с внешними базами знаний.

Предложена микросервисная архитектура, обеспечивающая интеграцию детерминированного и векторного поиска с промышленной масштабируемостью.

I. Архитектура платформы

Бэкенд реализован на Go (аутентификация, поиск) и Python/FastAPI (ML-компоненты). Микросервисы развёртываются в Docker/Kubernetes; взаимодействие через Apache Kafka; кэширование – Redis.

Обработка информации организована в три слоя:

Слой I — графовая семантическая память. Хранение верифицированных знаний в виде графов. Подключаются OSTIS [4] и Microsoft GraphRAG для логического вывода.

Слой II — поисковый микросервис. Embedding-модели генерируют векторы из документов. Хранение в PostgreSQL. Elasticsearch обеспечивает полнотек-

стовый поиск. Слияние результатов методом rank fusion.

Слой III — модуль генерации. Языковые модели синтезируют ответ по найденному контексту с контролем галлюцинаций.

МСП-агент позволяет LLM выполнять поисковые команды с правами доступа пользователя (RBAC).

II. Экспериментальная верификация

Пилотное исследование с участием двух компаний из Минска (50 сотрудников, 3 месяца) выявило эффективность: время ответа 0,5 с вместо 15 мин ручного поиска, снижение нагрузки на экспертов на 40%, сокращение времени поиска на 70–85%.

III. Выводы

Гибридная архитектура WikiAi обеспечивает интеграцию детерминированного и вероятностного поиска. Экспериментально подтверждена эффективность rank fusion (+28% nDCG@5). Результаты пилотного внедрения подтверждают сокращение операционных расходов.

Список литературы

1. McKinsey & Company. Solving the Productivity Puzzle, 2018.
2. IDC. FutureScape: Worldwide Digital Transformation 2020.
3. IDC Research. The Hidden Costs of Information Chaos, 2020.
4. GitHub-организация OSTIS. <https://github.com/ostis-ai>.

Афанасьев Иван Андреевич, учащийся 11 «А» класса ГУО «Гимназия №1 г. Бреста имени Защитников Брестской крепости».

Научный руководитель: Зотов Никита Владимирович, ассистент кафедры ИИТ БГУИР, магистр.

Научный руководитель: Горбачевич Ирина Юрьевна, учитель информатики ГУО «Гимназия №1 г. Бреста имени Защитников Брестской крепости».