Описание функциональных характеристик Arenadata Advanced RAG (ARAG)

Москва
2025

Содержание:

Термины и определения

Термин	Значение
ANSI SQL	Стандартный язык запросов к базам данных, разработанный Американским национальным стандартом (ANSI)
LLM-модели	Языковые модели, обученные на больших объемах текстовых данных
Открытая библиотека искусственного интеллекта	Набор алгоритмов, предназначенных для разработки технологических решений на основе искусственного интеллекта, описанных с использованием языков программирования и размещенных в информационно-телекоммуникационной сети "Интернет" (далее - сеть "Интернет")
Большие генеративные модели	Модели искусственного интеллекта, способные интерпретировать (предоставлять информацию на основании запросов, например об объектах на изображении или о проанализированном тексте) и создавать мультимодальные данные (тексты, изображения, видеоматериалы и тому подобное) на уровне, сопоставимом с результатами интеллектуальной деятельности человека или превосходящем их
Параметры модели искусственного интеллекта	Числовые значения, определяющие работу модели искусственного интеллекта, в частности выведение закономерностей, принятие решений или прогнозирование результатов

Сокращения и обозначения

Сокращение	Наименование
API	(англ. – Application Programming Interface) – набор классов, процедур, функций, структур или констант, которыми одна компьютерная программа может взаимодействовать с другой программой
CPU	(англ. – Central Processing Unit) – центральный процессор
IP	(англ. – Internet Protocol) – маршрутизируемый протокол сетевого уровня стека TCP/IP
LDAP	(англ. – Lightweight Directory Access Protocol) – протокол прикладного уровня для доступа к службе каталогов
LLMs	Large language models
MSSQL	Microsoft SQL Server

Сокращение	Наименование
SQL	(англ. – Structured Query Language) – декларативный язык программирования, применяемый для создания, модификации и управления данными в реляционных базах данных
SSL	Secure Sockets Layer
UI	User interface
БД	База данных
ГБ	Гигабайт
ИИ	Искусственный интеллект
ОС	Операционная система
ПО	Программное обеспечение
РФ	Российская Федерация
СУБД	Система управления базами данных
ТБ	Терабайт

1 Введение

Появление RAG-систем стало закономерным ответом на комплекс технологических вызовов и растущих требований к ИИ в реальных сценариях использования. Изначально языковые модели демонстрировали впечатляющие способности к генерации текста, но столкнулись с принципиальными ограничениями – они оперировали исключительно знаниями, зашитыми в их параметрах во время обучения, что приводило к двум критическим проблемам. Во-первых, модели часто выдавали убедительно звучащие, но фактически неверные ответы, когда сталкивались с темами за пределами их тренировочных данных, – этот феномен получил название «галлюцинаций». Во-вторых, их знания быстро устаревали, так как обновление весов модели требовало дорогостоящего переобучения. Параллельно в различных отраслях – от медицины до финансовых услуг – усиливались требования к прозрачности и проверяемости решений ИИ: пользователям недостаточно было получить ответ, им нужно было понимать, на каких источниках он основан, особенно в регулируемых областях с высокими требованиями.

Arenadata Advanced RAG (ARAG) – это продвинутая платформа смыслового поиска (Retrieval Augmented Generation), которая помогает ИИ и людям находить точные ответы на специфичные для компании вопросы с учетом внутренних данных, регламентов и специфики отрасли.

ARAG сочетает различные стратегии векторного поиска, обеспечивая максимальную полноту и точность извлечения знаний. Модульная архитектура позволяет выборочно применять наиболее подходящие инструменты обработки информации для групп документов, отдельных источников данных или целых доменов знаний, а также выбирать наиболее подходящие для задачи ИИ-модели. Платформа спроектирована для интеграции в корпоративные среды и масштабируется в зависимости от бизнес-сценариев.

2 Требования к программному обеспечению

Установка ARAG выполняется на хост под управлением Linux[1]. Для работы ARAG требуется наличие на хосте установленной и настроенной платформы запуска контейнерных приложений Docker или аналогичной, например, Podman. Текущему пользователю должны быть предоставлены привилегии запуска docker-контейнеров.

Системные требования для узла развертывания ARAG:

Ядро системы (Core Advanced RAG):

- CPU: 8+ ядер (Intel Xeon/AMD EPYC)
- RAM: 32+ ГБ
- Диск: 100+ ГБ SSD (для кэша и временных данных)
- Сеть: 1+ Гбит/с
- ОС: Linux (Ubuntu 22.04 LTS)
- Docker

Векторная база данных (Vector DB):

- CPU: 8+ ядер
- RAM: 64+ ГБ
- Диск: 1+ ТБ NVMe SSD
- Отдельная нода (рекомендуется)
- Docker

Графовая база данных (Graph DB):

- CPU: 8+ ядер
- RAM: 64+ ГБ
- Диск: 500+ ГБ SSD
- Отдельная нода (рекомендуется)
- Docker

[1] Работа ПО на Windows возможна, но не описывается в данной инструкции.

Требования к LLM серверу:

- GPU: 2+ x NVIDIA 6000Ada (48+ ГБ VRAM)
- / 2+ x NVIDIA 4090/3090 (24+ ГБ VRAM)
- CPU: 16+ ядер
- RAM: 64+ ГБ
- Диск: 500+ ГБ SSD
- Отдельный GPU-кластер (обязательно)
- Docker

Требования к Embedding-серверу:

- GPU: 1 x NVIDIA 4090/3090 (24+ ГБ VRAM)
- CPU: 8+ ядер
- RAM: 64+ ГБ
- Диск: 500+ ГБ SSD
- Отдельный GPU-кластер (не обязательно)
- Docker

Требования к RAGAS (Evaluation module)

- CPU: 4+ ядер
- RAM: 32+ ГБ
- Диск: 100+ ГБ (для тестовых датасетов)
- Docker

ОС: CentOS 7.9 / Ubuntu / Astra Linux;
Системное ПО: Docker / Docker CE / Podman.

Требования к сетевой инфраструктуре

Внутренние подключения: сетевая связанность между узлами решения, отсутствие блокировок портов (TCP / UDP), производительность не ниже 1 Gb / sec.

3 Функциональные характеристики

Основные функции продукта:

- Семантический поиск – находит текст, близкий по смыслу к запросу, даже если нет точного совпадения ключевых слов;
- Поиск релевантных документов – извлекает информацию из внешних источников;
- Фильтрация и ранжирование – отбирает наиболее подходящие фрагменты текста по релевантности;
- Поддержка разных форматов – работает с файлами разных форматов, имеет возможность интеграции с источниками данных по API;
- Контекстно-зависимые ответы – использует найденные документы для генерации точных и развернутых ответов;
- Обобщение информации – может суммировать длинные документы или несколько источников;
- Ссылки на источники – сохраняет мета-данные исходного документа;
- Мультиязычность – поддерживает поиск и генерацию на разных языках.

При этом к функциям предъявляются следующие требования:

1. Поддерживаемые форматы файлов: PDF (включая таблицы), DOCX, XLSX, HTML, Markdown, TXT, CSV и другие.
2. Интеграция с одним из самых распространенных вики-решений.
3. Поддержка доменов при загрузке данных.
4. Использование Raptor, DenseX и LigthRAG подходов для формирования эмбеддингов и последующего поиска.
5. Поиск по графам знаний.
6. Технический GUI.

7. Наличие модуля оценки качества поиска – RAGAS.
8. API для загрузки данных и поиска.
9. Поддержка локальных моделей:

- Эмбеддинги: all-MiniLM-L6-v2, BGE-m3
- LLM: Qwen3-32B
- Реранкер: colbertv2.0

Поддерживаемые источники документов:

- Загрузка пространств самого распространенного вики-решения.
- Локальная папка с файлами.

В параметре loader.selected_loader конфигурации указывается источник confluence_loader или simple_directory_reader.

Загрузка документов может быть осуществлена двумя способами:

- через UI;
- curl.

При загрузке новых документов указывается домен, в который они будут загружены.

Поддержка загрузки Confluence spaces:

- Поддержка конфигурации со списком пространств;
- Возможность указания списка пространств в конфигурационном файле. При наличии списка загружаются только указанные пространства;
- Массовая загрузка пространств;
- Возможность получения всех пространств через API-запрос;
- Итеративная загрузка всех доступных пространств и соответствующих страниц/данных.

Поддержка логики обновления:

- Идентификация обновленного документа:
    - список id страниц, измененных за последние сутки статей
    - список id страниц, вновь созданных за последние сутки статей
    - список id страниц, удаленных за последние сутки статей
- Поиск и удаление связанных чанков из векторной БД (маппятся по id статей id родительских чанков и соответственно id дочерних чанков. Получаем полный список всех чанков, по которым произошли изменения за последние сутки)
- Повторная обработка и загрузка документа (только для метода DenseX старые чанки удаляются, формируются новые чанки по id страниц).

Все ошибки загрузки фиксируются стандартным способом (через текущую систему логирования), без дополнительных механизмов обработки.

В случае падения Docker-контейнера или критической ошибки, после восстановления система не продолжает загрузку.

Восстановление выполняется вручную: оператор проверяет последнее успешно загруженное пространство и вносит нужный список в конфиг.

В случае ошибки при загрузке пространства или страницы, они пропускаются, и начинается обработка следующего доступного объекта.

Должны поддерживаться следующие параметры:

- confluence_url: URL-адрес пространства Confluence, например, https://templates.atlassian.net/wiki.
- user_name: Имя пользователя Confluence.

- api_token: токен API, который можно получить здесь https://id.atlassian.com/manage-profile/security/api-tokens .
- space_key: идентификатор пространства Confluence.
- limit: ограничение на количество загружаемых страниц.
- page_ids: Список идентификаторов страниц.

Загрузка из локальной папки

При загрузке из локальной папки должны поддерживаться следующие параметры:

- input_dir : Каталог, содержащий данные.
- recursive: Указывает, следует ли выполнять рекурсивный обход каталога.
- docling :

o enabled: Включает или отключает использование Docling для файлов PDF. Установите true для активации.

o device : Указывает устройство для запуска вывода модели. Доступные параметры:

- "auto"– автоматический выбор устройства.
- "cpu"– использовать ЦП.
- "cuda"– использовать графический процессор с CUDA.

o num_threads: Количество потоков, используемых во время обработки.

3.1 Интерфейсы взаимодействия

ПО ARAG предоставляет пользовательский интерфейс администратора (UI).

Техническая консоль позволяет выполнять как загрузку, так и поиск.

При загрузке новых документов, они загружаются в указанный домен. При поиске – поиск будет осуществлен в рамках указанного домена. Если домен не указан, поиск осуществляется по домену, найденному с помощью алгоритма на основе ключевых слов.

При поиске на панели Pipeline Stages отображается техническая информация о всех метаданных о чанках.