УДК 004.08

Афанасьева Мария Сергеевна,
бакалавр,
Российский государственный университет нефти и газа (НИУ) имени И.М. Губкина,
Г. Москва

РАЗРАБОТКА КОГНИТИВНОГО АССИСТЕНТА ОПЕРАТОРА ЦЕНТРА УПРАВЛЕНИЯ ПРОИЗВОДСТВОМ

Аннотация. *Статья посвящена разработке когнитивного ассистента оператора НПЗ на основе RAG-подхода. Проанализирована деятельность оператора, выявлены недостатки существующих систем. Представлены архитектура ассистента, база знаний из технической документации, модели эмбеддингов и LLM. Разработана библиотека для индексации, поиска и генерации ответов. Тестирование подтвердило высокую релевантность поиска и корректную работу при отсутствии информации.*

Ключевые слова: *когнитивный ассистент, RAG, оператор ЦУП, нефтепереработка, нейросетевые технологии, большие языковые модели, эмбеддинги, семантический поиск, база знаний.*

DEVELOPMENT OF A COGNITIVE ASSISTANT FOR PRODUCTION CONTROL CENTER OPERATOR

Abstract. *The article presents the development of a cognitive assistant for a refinery control center operator based on the RAG approach. The operator's activities are analyzed, and shortcomings of existing systems are identified. The assistant architecture, technical documentation knowledge base, embedding and LLM models are described. A library for indexing, search, and response generation*

is developed. Testing confirmed high search relevance and correct behavior when information is missing.

Keywords: *cognitive assistant, RAG, control center operator, oil refining, neural network technologies, large language models, embeddings, semantic search, knowledge base.*

Эффективное управление производственными процессами зависит от оперативности работы персонала центров управления производством (ЦУП). В любой момент оператору ЦУП может потребоваться обратиться к технической документации. От того, насколько быстро он найдет информацию, зависит скорость принятия решений и безопасность объекта.

Традиционные подходы к поиску информации имеют ряд недостатков: они не учитывают семантику запроса, не понимают контекст и не всегда выводят релевантные фрагменты. Существующие системы электронного документооборота (СЭД) обеспечивают централизованное хранение файлов, но их поиск ограничен ключевыми словами. Корпоративные решения, такие как 1С:Документооборот [1] и SAP, не поддерживают обработку вопросов на естественном языке, а их внедрение требует значительных затрат.

После выявления недостатков существующих технологий принято решение разработать собственного когнитивного ассистента оператора ЦУП. Такой подход сочетает использование открытых нейросетевых моделей, локальное развертывание для безопасности данных и прозрачность ответов за счет указания источников.

Цель работы – разработка когнитивного ассистента оператора ЦУП на основе нейросетевых технологий и базы знаний, обеспечивающего поиск релевантной информации в технической документации и формирование ответа на естественном языке с указанием источника, что позволяет сократить время поиска, снизить нагрузку на персонал и повысить надежность решений.

ЦУП представляет собой сложную многоуровневую систему, объединяющую технологические объекты, интеллектуальные АСУ ТП, системы для сбора и обработки данных, а также программные комплексы поддержки принятие решений [2]. Вся описанная инфраструктура, от датчиков до видеостены, работает как единое целое, обеспечивая эффективное управление производством. Данная архитектура получила название цифровой экосистемы предприятия, иерархия которой представлена на рисунке 1.



Рисунок 1. Цифровая экосистема производства

На основе проведенного при выборе подхода к проектированию сравнительного анализа, наиболее применимым для большинства корпоративных задач является RAG-подход [3].

Работа RAG-системы делится на два этапа: подготовка базы знаний и обработка запроса [4]. База знаний наполняется технической литературой и производственной документацией. На этапе подготовки документы разбиваются на чанки, каждый преобразуется в вектор-эмбеддинг фиксированного размера [5]; векторы сохраняются с метаданными в векторном хранилище. При запросе система преобразует вопрос в вектор, который передаётся ретриверу, находящему релевантные чанки по

косинусному расстоянию [6]. Косинусное сходство и расстояние вычисляются по формулам (1) и (2). Косинусное сходство:

$$\cos_sim(q,d) = \frac{q \cdot d}{||q|| * ||d||} = \frac{\sum_{i=1}^{n} q_i \cdot d_i}{\sqrt{\sum_{i=1}^{n} q_i^2} \cdot \sqrt{\sum_{i=1}^{n} d_i^2}} \quad (1)$$

Косинусное расстояние:

$$\cos_dist(q,d) = 1 - \cos_sim(q,d) \quad (2)$$

где q – вектор вопроса;

d – вектор чанка.

Таким образом, главная задача ретривера состоит в том, чтобы подобрать такие чанки из векторного хранилища, косинусное расстояние которых с вектором вопроса будет минимальным.

После нахождения наиболее релевантных данных, системой формируется расширенный контекст, он передается в LLM, которая, объединив полученную информацию с системным промптом, формирует ответ и возвращает его в пользовательский интерфейс.

Разработка выполнена на Python благодаря его экосистеме для машинного обучения, кроссплатформенности и простоте. Для запуска языковой модели используется Ollama, обеспечивающий локальную работу и REST API. Выбрана модель эмбеддингов paraphrase-multilingual-MiniLM-L12-v2. Для генерации используется Qwen2.5:1.5b. В стек также входят PyPDF для извлечения текста и ChromaDB для семантического поиска.

Первым этапом тестирования стала проверка корректности индексации документов. Логирование не содержало критических ошибок. В консоль выводилась детальная информация о процессе, отраженная на рисунке 2.

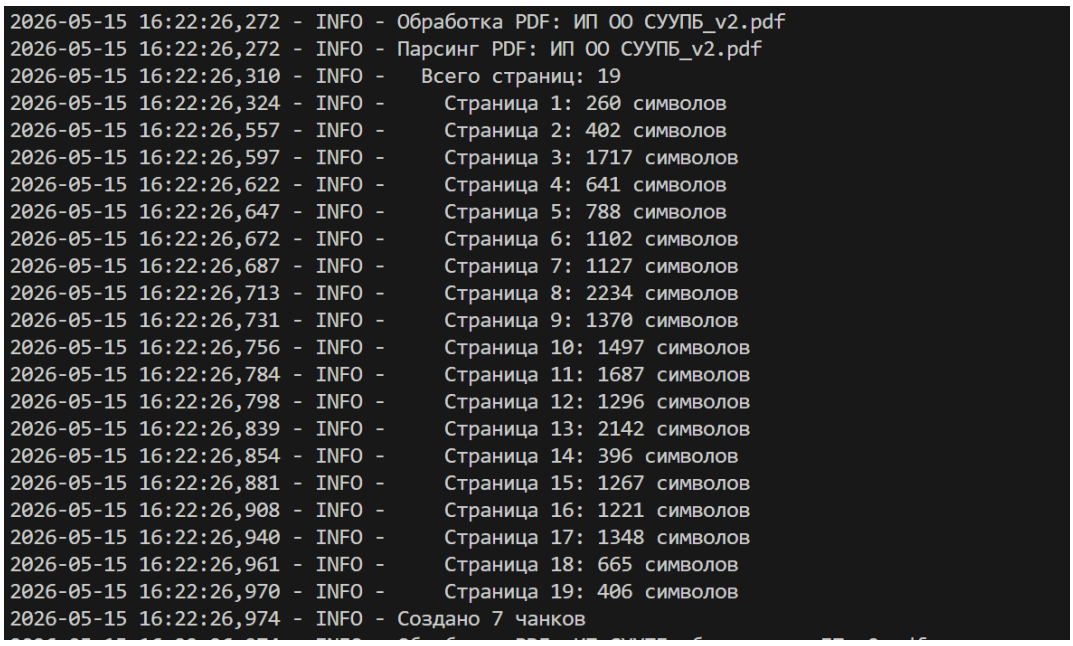


Рисунок 2. Результат загрузки полученных чанков в векторное хранилище на примере одного документа

При старте программы в консоли вывелось приветственное сообщение с краткой справкой о доступных командах управления сессией, после чего появилось приглашение для ввода вопроса оператора. Результат запуска ассистента представлен на рисунках 3 и 4.

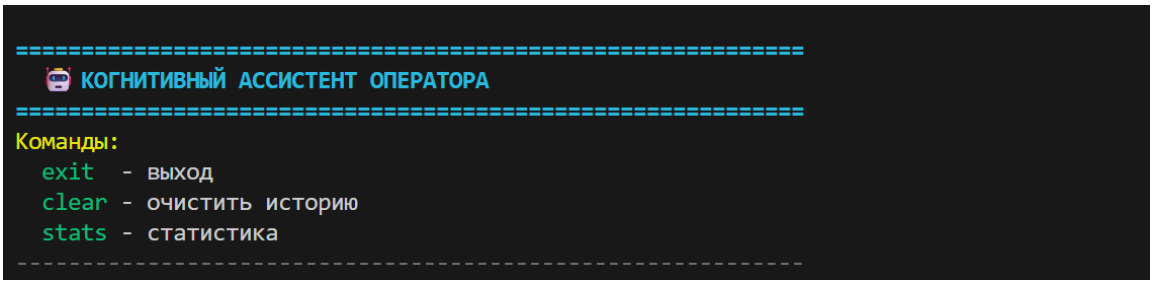


Рисунок 3. Стартовый текст



Рисунок 4. Предложение задать вопрос

После успешного запуска ассистента необходимо проверить сематический поиск, а именно: способность разграничивать вопросы,

относящиеся к теме документации, и вопросы, ответы на которые в документации отсутствуют.

Для этого в рамках одной сессии были последовательно заданы два вопроса.

Первый тестовый вопрос «как отключить модуль управления смешением?» уместен, так как оператор имеет доступ к документации, и ассистент должен найти ответ в базе знаний (рис. 5).

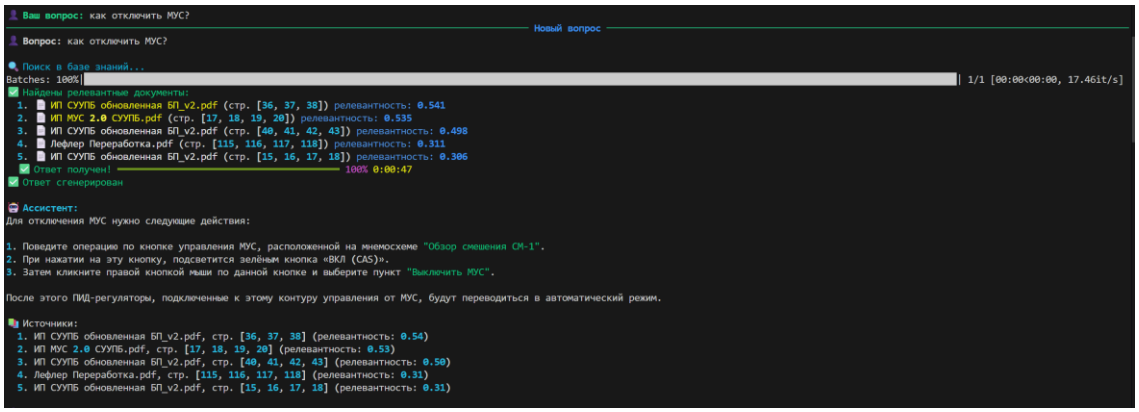


Рисунок 5. Ответ ассистента на первый тестовый вопрос

На рисунке видно, как ассистент выполнил семантический поиск, нашёл релевантные фрагменты с указанием релевантности и сгенерировал ответ по найденному контексту из источников.

В качестве второго вопроса, ответа на который, очевидно, в документации нет, была просьба написать рецепт для приготовления блинчиков. Результат тестирования представлен на рисунке 6.

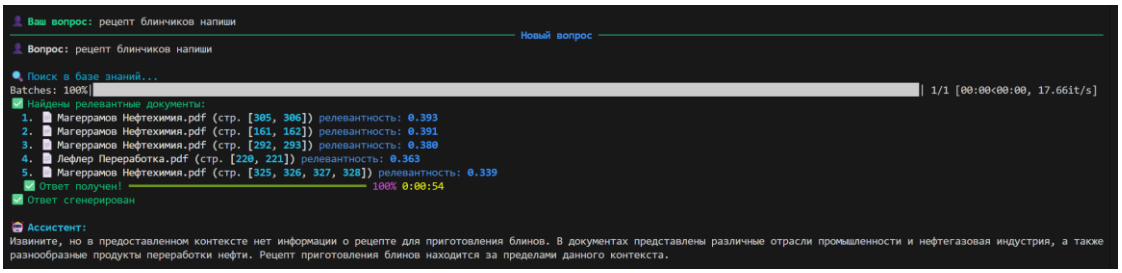


Рисунок 6. Ответ ассистента на второй тестовый вопрос

Несмотря на то, что ретривер обнаружил в документации релевантные чанки, вопрос не касался деятельности оператора, поэтому ассистент не стал «выдумывать» ответ и уведомил пользователя об отсутствии соответствующей информации.

На предложенной ассистентом странице находился параграф о свойствах масел – семантический поиск зафиксировал близость между понятиями «блины» и «масло» (рис. 7).



25. 4. Низкотемпературные свойства масел

Ко многим сортам смазочных масел предъявляются жёсткие требования по их прокачиваемости при низких температурах. Масла для холодильных машин, приборные, моторные и некоторые другие должны по условиям эксплуатации не терять подвижности при температурах от -30 до -60 ^{0}С. В технических нормах это качество масла контролируется определением их температуры застывания.

Рисунок 7. Информация, содержащаяся на предложенной ассистентом странице документа

В качестве количественной оценки эффективности было измерено время ответа ассистента на четыре тестовых запроса. Для каждого вопроса, кроме просьбы написать рецепт, фиксировалось полное время от отправки сообщения до вывода сгенерированного ответа. На основе полученных значений по формуле (3) было рассчитано среднее время ответа ассистента:

$$t_{avg} = \frac{1}{n}\sum_{i=1}^{n} t_i \,, \qquad (3)$$

$$t_{avg} = \frac{1}{4}(47 + 64 + 74 + 54) \approx 60 \text{ сек.} \approx 1 \text{ мин.}$$

Среднее время ответа ассистента вдвое ниже критического порога в 120 секунд для нештатных ситуаций, что подтверждает соответствие нормативам оперативной работы при точных ссылках на источники.

Таким образом, разработанный когнитивный ассистент оператора ЦУП позволяет получать ответы на естественном языке по документации с

указанием источников, сокращая время поиска. Тестирование подтвердило корректность ответов на технические вопросы и отклонение запросов вне предметной области, что позволяет рекомендовать его к опытно-промышленной эксплуатации.

Литература:

1. Павлов, В. Обзор возможностей 1С:Документооборот 8 и описание конфигурации [Электронный ресурс] // wiseadvice-it.ru. – URL: https://wiseadvice-it.ru/o-kompanii/blog/articles/1s-dokumentooborot-8-obzor-opisanie-vozmozhnosti/ (дата обращения: 16.05.2026).
2. Цифровой завод: что уже сделано на Московском и Омском НПЗ [Электронный ресурс] // up-pro.ru. – URL: https://up-pro.ru/library/information_systems/automation_production/tsifrovoy-zavod/ (дата обращения: 06.05.2026).
3. Добренко, Н. В., Добренко, Д. А., Улизько, М. В. Интеллектуальная поддержка принятия управленческих решений в MES-системах с использованием больших языковых моделей // Экономика. Право. Инновации. – 2024. – № 3. – С. 47–59.
4. Инкижеков, А. А. Улучшение генерации текста с помощью Retrieval Augmented Generation // Инженерные технологии: традиции, инновации, векторы развития : материалы X Всероссийской научно-практической конференции с международным участием, Абакан, 13–15 ноября 2024 года. – Абакан : Хакасский государственный университет им. Н. Ф. Катанова, 2024. – С. 27–29.
5. Ли Цзымин, Ковальчук, Е. А., Мишра, А., Потехин, В. В. Интерактивный метод диагностики неисправностей на основе больших языковых моделей [Электронный ресурс] // Промышленный искусственный интеллект (ПИИ’2025) : сборник научных трудов Всероссийской научно-практической конференции с международным участием, Санкт-Петербург, 3–

4 июля 2025 года. – Санкт-Петербург : ПОЛИТЕХ-ПРЕСС, 2025. – URL: https://elib.spbstu.ru/dl/2/id25-526.pdf (дата обращения: 01.05.2026).

6. Косинусная близость [Электронный ресурс] // textarget.ru. – URL: https://textarget.ru/encyclopedia/kosinusnaja-blizost/ (дата обращения: 16.05.2026).