УДК 004.891
DOI 10.52684/2312-3702-2024-50-4-97-101

ИНТЕЛЛЕКТУАЛИЗАЦИЯ ПРОЦЕССА АДАПТАЦИИ НОВЫХ СОТРУДНИКОВ В СТРОИТЕЛЬНЫХ КОМПАНИЯХ

А. С. Панкрашов, С. В. Окладникова

Панкрашов Александр Сергеевич, ведущий инженер-разработчик отдела внедрения информационных систем, ООО «Алиал Групп», г. Санкт-Петербург, Российская Федерация; e-mail: a.pankrashov@alial.group;

Окладникова Светлана Владимировна, кандидат технических наук, доцент кафедры систем автоматического проектирования и моделирования, Астраханский государственный архитектурно-строительный университет, г. Астрахань, Российская Федерация; e-mail: okladnikova.s.v@yandex.ru

Цифровая трансформация предусматривает внедрение искусственного интеллекта в различные бизнес-процессы компании, включая онбординг, который позволяет новым сотрудникам быстрее влиться в команду. В 2024 году нехватка кадров в строительной сфере составляет около 15 %. Одной из причин является неправильное выстраивание этапов подбора и адаптации персонала. Поэтому около 50 % сотрудников меняют место работы в течение полугода с момента найма. Использование технологий искусственного интеллекта в онбординг позволит компаниям повысить свою конкурентоспособность за счет снижения текучести кадров, а новым сотрудникам быстрее адаптироваться к новым условиям труда. В статье представлены результаты исследований, проведенные авторами, по возможности применения больших языковых моделей и RAG-архитектуры для разработки экспертной системы по поддержке процесса онбординга.

Ключевые слова: *онбординг, экспертная система, большая языковая модель, RAG-архитектура, база знаний, искусственный интеллект, промпт, эмбеддинг, база знаний, векторное хранилище.*

INTELLECTUALIZATION OF THE PROCESS OF ADAPTATION OF NEW EMPLOYEES IN CONSTRUCTION COMPANIES

A. S. Pankrashov, S. V. Okladnikova

Pankrashov Aleksandr Sergeyevich, Leading Development Engineer of the Information Systems Implementation Department, Alial Group LLC, Saint Petersburg, Russian Federation; e-mail: a.pankrashov@alial.group;

Okladnikova Svetlana Vladimirovna, Candidate of Technical Sciences, Associate Professor of Automatic Design and Modeling Systems Department, Astrakhan State University of Architecture and Civil Engineering, Astrakhan, Russian Federation; e-mail: okladnikova.s.v@yandex.ru

Digital transformation involves the introduction of artificial intelligence into various business processes of the company, including onboarding, which allows new employees to integrate into the team faster. In 2024, the shortage of personnel in the construction sector is about 15%. One of the reasons is the incorrect alignment of the stages of recruitment and adaptation of personnel. Therefore, about 50% of employees change their place of work within six months from the moment of hiring. The use of artificial intelligence technologies in onboarding will allow companies to increase their competitiveness by reducing staff turnover, and new employees will adapt faster to new working conditions. The article presents the results of research conducted by the authors on the possibility of using Large Language Models and RAG architecture to develop an expert system to support the onboarding process.

Keywords: *onboarding, expert system, large language model, RAG architecture, knowledge base, artificial intelligence, industrial, embedding, knowledge base, vector storage.*

Введение

В настоящее время существуют успешные практики применения технологий искусственного интеллекта (Artificial Intelligence, Blockchain, Big Data и др.) при проектировании и строительстве различных объектов, которые генерируют и принимают конкретные решения, составляют перспективные прогнозы на заданный срок, диагностируют и решают проблемы с минимальным участием человека в данных процессах [1]. Используемая в строительной индустрии методология BIM-моделирования (Building Information Modeling) поддерживает в первую очередь технологические процессы. Она ориентирована на работу с информационными 3D-моделями зданий и сооружений на всех этапах жизненного цикла строительного объекта [2].

По данным интернет-рекрутмента строительная отрасль входит в топ-5 отраслей с самым высоким кадровым дефицитом, который в 2023 году вырос на 31 %. Причины возникшего дефицита связаны с увеличением объема строительства, сокращением числа квалифицирован-

ных специалистов, а также плохой организацией HR-процессов внутри компании, в частности процесса адаптации нового сотрудника (онбординг) к новым условиям работы, которая во многом определяется физиологическими и психологическими особенностями человека и существующими социальными проблемами [3].

Около 50 % российских компаний с целью быстрого погружения новых сотрудников в рабочие процессы внедряют программы онбординга. Чем более гибкая и эффективная их организация, тем больше шансов, что новый сотрудник останется работать в компании, соответственно меньшие потери будет нести компания [4]. Проведенные в 2023 году Skillbox исследования показали, что у 44 % компаний отсутствуют ресурсы для разработки системы онбординга, 22 % – не видят ее значимости, 17 % – не понимают, как нужно ее разрабатывать, а 5 % – вообще не знакомы с этим понятием. Проблемы онбординга связаны в первую очередь с отсутствием в компании автоматизации и цифровых инструментов взаимодействия с новыми сотрудниками, что не позволяет выстроить с ними необходимую коммуникацию, хаотично работающими системами наставничества, обучения и продвижения, отсутствием или недостатком обучающих материалов и в целом базы знаний [5].

Сегодня многие компании для улучшения своих HR-процессов применяют различные цифровые инструменты, основанные на технологиях искусственного интеллекта (ИИ): экспертные системы, персонализированные ассистенты, чат-боты, системы мониторинга прогресса новых сотрудников и др. Их использование способствует повышению эффективности кадровых процессов на 30 %. Технологии ИИ позволяют перейти к информационным системам HR-менеджмента на естественном языке, образуя единую среду для проектирования систем LLM-агентов, общающихся с программным обеспечением на программном уровне, а с человеком – на естественном языке [6, 7].

В статье рассмотрены современные подходы реализации архитектуры RAG и больших языковых моделей (LLM) [8], а также особенности их использования для разработки экспертных систем (ЭС), обеспечивающих поддержку HR-процессов.

Методы

В настоящее время одной из альтернатив классической архитектуры экспертных систем выступают LLM, которые обладают большим потенциалом к внедрению в качестве машины логического вывода как основного логического инструмента ЭС. Они полностью ориентированны на анализ контекстуальных связей в тексте и демонстрируют высокие показатели по скорости обработки информации [9, 10].

Расширенная поисковая генерация (RAG) – это подход, который использует контекст LLM с целью расширения имеющихся в самой модели знаний. Под контекстом понимается размерность текста, который большая языковая модель может обработать единовременно. Техники разработки и проектирования запросов (промптов), передаваемых на обработку большой языковой модели, ориентированы на решение конкретных задач. Существуют как более простые, так комплексные подходы к их реализации (Zero-shot, Chain-of-Thought и др.) [11–13]. Если передать языковой модели вопрос и участок информации, в котором содержится ответ на него, то она сможет успешно дать правильный ответ без искажений. На рисунке 1 представлена схема архитектуры RAG-системы.

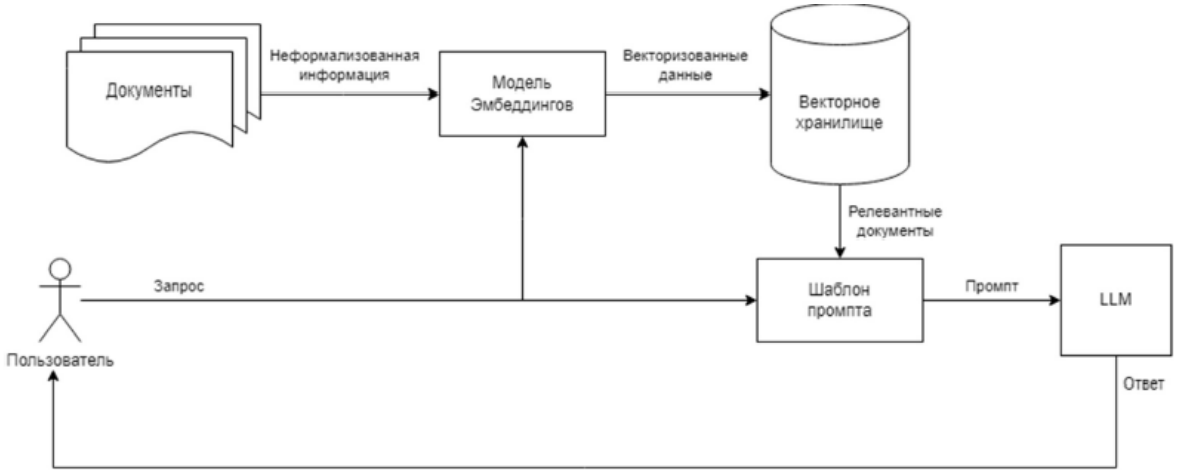

Рис 1. Архитектура RAG-системы

Стилевые и контекстуальные особенности общения с большой языковой моделью обусловлены ее особенностью при обучении выстраивать вероятные последовательности слов, семантически значимые, несущие смысловую нагрузку и отвечающие требованиям переданных инструкций

[14, 15]. Готовые инструкции переводят в «машиночитаемый» вид с помощью алгоритмов эмбеддинга, то есть приводят «человекочитаемый» текст в набор чисел – вектор. Для этого выполняют последовательное деление цельного текста на более мелкие единицы [16]:

1) корпусы – группы документов, проходящих процедуру векторизации;

2) документы – группы текста, которые несут отдельную смысловую нагрузку (параграф учебника, научная статья и т. п.);

3) токены – являются минимальной текстовой единицей, представляют собой слова или словосочетания.

Результатом поэтапной декомпозиции документов становится набор коллекций с определенным уровнем вложенности, зависящим от характера процедуры деления на составные элементы.

Следующий этап – процесс преобразования данных на естественном языке в векторные представления. Наиболее простым является подход «Прямого кодирования» («One-hot encoding»), который реализуется с учетом следующих условий [17, 18]:

1) токен представляет собой бинарный вектор;

2) все токены группируются в общий словарь (который может быть отсортирован по алфавиту);

3) вектором конкретного слова станет последовательность чисел, где элемент вектора, соответствующий позиции слова в словаре, принимает значение, равное 1.

В современных моделях векторизации текстов размеры конечных векторов достигают 1024 значений. Для их обработки существует отдельный вид баз данных (БД) – векторные хранилища. Их функциональные возможности включают в себя не только хранение, но и поиск данных. В отличие от классических систем управления базами данных (СУБД) в векторных хранилищах применяются специальные алгоритмы, которые производят поиск по признакам подобия векторов. Каждое векторное хранилище работает с разными алгоритмами поиска, поэтому при выборе подходящего инструмента критериями являются производительность и эффективность [19, 20].

Результаты и обсуждение

Учитывая особенности применения LLM и построения систем на основе RAG-архитектуры, можно рассматривать их как аналоги классической архитектуре экспертных систем, то есть машину логического вывода в экспертной системе заменяет большая языковая модель, а базу знаний – векторное хранилище. Дополнительных алгоритмов по приведению запросов на естественном языке к машиночитаемому виду для самой языковой модели не требуется, а алгоритмы эмбеддингов существуют в разных вариациях и могут обеспечить различный уровень качества.

Таким образом, можно сделать вывод, что, заменив экспертную систему RAG-системой, целевой потребитель не только получает эксперта с определенным набором знаний, но и гибкое решение, в котором компоненты свободно заменяются и переиспользуются. Существующие техники промпт-инжиниринга позволяют без особых усилий частично изменить результаты, не прибегая к программному коду.

В рамках данного подхода присутствуют определенные недостатки. Производительность генерации большой языковой модели на CPU (центральный процессор) крайне мала. Для RAG-системы целесообразнее использовать GPU (графический процессор), который поддерживает Cuda Toolkit с целью перенести все вычисления на графическое устройство [21].

Для анализа производительности большой языковой модели при решении задач генерации текста на GPU или на CPU из датасета «IlyaGusev/ru_sharegpt_cleaned», размещенного на портале HuggingFace, было выбрано пять случайных запросов. Тестирование проводилось при помощи программы LM Studio, в качестве LLM – «TheBloke/saiga_mistral_7b-GGUF», версия с 6-битной квантизацией. Как CPU использовался процессор Intel Core i5-12400F, как GPU – NVidia RTX 3060 8 GB. Графическая иллюстрация полученных в ходе эксперимента результатов приведена на рисунке 2. Для всех пяти запросов обычный процессор занимал существенно больше времени, чем графическое ядро. В случае, если итоговая RAG-система недостаточно качественно отвечает на вопросы, можно расширить знания модели за счет процедуры дообучения (fine tuning). Проблемой полного переобучения большой языковой модели является время, которое представляет собой один из ценных ресурсов для любого предприятия. Поэтому актуальным является выбор в пользу более быстродействующего решения.

В системах с поддержкой логического вывода быстродействие должно соответствовать качеству вывода с точки зрения полноты информации. На сегодняшний день для вычислений в системах искусственного интеллекта используется программное обеспечение CUDA [22, 23]. Любые вычисления в области линейной алгебры, тем более выполняющиеся постоянно, быстрее всего осуществляются на базе GPU как самого эффективного и доступного на рынке решения. Соответственно реализация ЭС на базе вычислений за счет центрального процессора ведет к трате колоссального количества времени, что для любого производства губительно, так как лишает предприятие конкурентоспособности.

В ходе опытных испытаний авторами были выявлены следующие положительные аспекты

в пользу применения LLM при разработке экспертной системы для поддержки онбординга:

1) эффективность обработки естественного языка без необходимости самостоятельно обрабатывать информацию, поступающую от пользователя;

2) возможность расширить доступные знания модели собственными базами знаний в виде векторных хранилищ за счет применения архитектуры RAG.

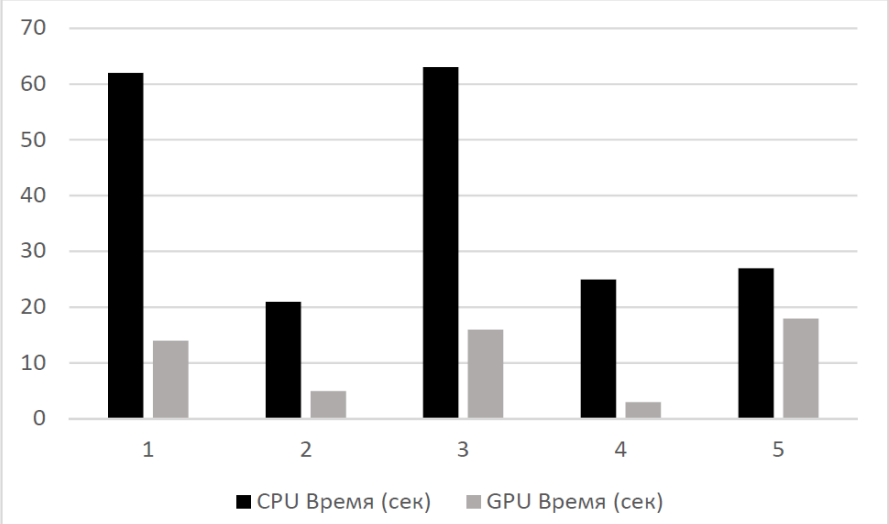

Рис. 2. Сравнение GPU и CPU в задачах генерации

За время проведения экспериментов было разработано пять решений, три из которых оказались концептуально схожими с программным обеспечением, разработанным для проведения исследования. Остальные два были направлены на интеллектуальную обработку документов и подготавливались командой, состоящей из разработчиков и системных аналитиков. Сбор информации выполнялся из электронного справочного ресурса компании Directum. Был разработан пользовательский сценарий и произведен парсинг статей справки. Из векторизованных текстов была сформирована отдельная база знаний (представляющая собой векторное хранилище).

Данное решение на базе большой языковой модели прошло тестирование на этапе пресейлов. Когда аналитики не могли дать точный ответ, они обращались к чат-боту и успешно отвечали на вопросы клиента по функционалу системы Directum RX и ее архитектуре.

В ходе исследования применения LLM и опытной эксплуатации были выявлены следующие особенности RAG-систем:

1) гибкость в эксплуатации – языковая модель успешно анализирует запрос и за счет собственных механизмов на уровне архитектуры генерирует ответы, которые в полной мере соответствуют требуемой задаче. Таким образом, ЭС на базе данной концепции способна оперировать несколькими областями знаний сразу, давая развернутый ответ и учитывая все возможные или заданные условия;

2) масштабные возможности – количество инструментов, агрегирующих программные компоненты и информационные ресурсы, на данный момент растет, а все имеющиеся решения получают стабильные обновления и успешно используется в производственной деятельности как отечественных, так и зарубежных компаний.

Таким образом, на данный момент большие языковые модели и RAG-системы являются одним из лучших вариантов для реализации экспертных систем. Для работы LLM необходимы видеокарты от NVidia, желательно серии RTX с целью использования инструментов CUDA. Данные GPU являются крайне дорогими, не каждый может позволить себе более одного такого устройства. В остальном, на фоне классических ЭС, RAG выигрывает в гибкости, скорости и многозадачности.

Заключение

Авторы представили основные результаты проводимого исследования по выбору альтернативного подхода к разработке экспертных систем для поддержки HR-процессов, в частности, онбординга, на основе инструментов искусственного интеллекта.

При помощи большой языковой модели (в качестве машины логического вывода) и векторного хранилища (в качестве базы знаний) был спроектирован и разработан прототип ЭС поддержки HR-процесса онбординга персонала в строительной компании. Разработка показала положительные результаты, в два раза снизив скорость поиска ответов на возникающие у новых сотрудников вопросы. Правильность ответов в среднем составила ~ 80 %.

Список литературы

1. Рыбкина Г. В. Промышленная цифровизация в строительстве: многоаспектный подход и ключевые технологии / Г. В. Рыбкина, И. А. Зайцева, С.А. Логинова, А. В. Симагин // Инженерно-строительный вестник Прикаспия. – 2024. – № 2 (48). – С. 77–84.
2. Байорис А. Р. BIM-моделирование как технология повышения конкурентоспособности строительной отрасли/ А. Р. Байорис, М. А. Малиновский, А. В. Ершов// Регулирование земельно-имущественных отношений в России: правовое и геопространственное обеспечение, оценка недвижимости, экология, технологические решения. – 2021. – № 1. – С. 79–83. – DOI 10.33764/2687-041X-2021-1-79-83.
3. Сафина Г. Л. Исследование рынка труда в строительной сфере / Г. Л. Сафина, Ю. И. Казяба // Инженерно-строительный вестник Прикаспия. – 2023. – № 1 (43). – С. 78–83.
4. Стройка в «цифре» // Комплекс градостроительной политики и строительства города Москвы. – Режим доступа: https://stroi.mos.ru/interviews/stroika-v-tsifrie (дата обращения: 16.09.2024), свободный. – Заглавие с экрана. – Яз. рус.
5. Водопьянова Н. А. Онбординг сотрудников как ключевой элемент в процессе адаптации персонала организации / Н. А. Водопьянова, Е. В. Баргатинова // Вестник Академии знаний. – 2024. – № 3 (62). – С. 716–719.
6. Искусственный интеллект в строительстве. Примеры ИИ для строительной отрасли // Деловая сеть. – Режим доступа: https://elport.ru/articles/iskusstvennyiy_intellekt_v_stroitelstve_primeryi_ii_dlya_stroitelnoy_otrasli (дата обращения: 02.09.2024), свободный. – Заглавие с экрана. – Яз. рус.
7. Окладникова С. В. Применение технологий искусственного интеллекта в HR-менеджменте / С. В. Окладникова, А. С. Панкрашов // Вестник Дагестанского государственного технического университета. Технические науки. – 2023. – № 50 (2). – С. 117–125.
8. Rag Architecture For Ai System Design // Restack. – Режим доступа: https://www.restack.io/p/ retrieval-augmented-generation-answer-rag-architecture-ai-system-design-cat-ai (дата обращения 10.09.2024), свободный. – Заглавие с экрана. – Яз. рус.
9. How RAG Architecture Overcomes LLM Limitations. Narendran N // The New Stack. – Режим доступа: https://thenewstack.io/how-rag-architecture-overcomes-llm-limitations/ (дата обращения 15.09.2024), свободный. – Заглавие с экрана. – Яз. рус.
10. Кузнецов А. В. Цифровая история и искусственный интеллект: перспективы и риски применения больших языковых моделей / А. В. Кузнецов // Новые информационные технологии в образовании и науке. – 2022. – № 5. – С. 53–57. – DOI 10.17853/2587-6910-2022-05-53-57.
11. Zero-Shot Prompting // Prompt Engineering Guide. – Режим доступа: https://www.promptingguide.ai/ru/techniques/zeroshot (дата обращения: 23.09.2024), свободный. – Заглавие с экрана. – Яз. рус.
12. Руководство по промпт-инжинирингу // Prompt Engineering Guide. – Режим доступа: https://www.promptingguide.ai/ru (дата обращения: 22.09.2024), свободный. – Заглавие с экрана. – Яз. рус.
13. A Guide to LLM Hyperparameters // symbl.ai. – Режим доступа: https://symbl.ai/developers/ blog/a-guide-to-llm-hyperparameters/ (дата обращения: 12.07.2024), свободный. – Заглавие с экрана. – Яз. рус.
14. Rag Architecture Easy Explained // DevArt. – Режим доступа: https://dev-art.vercel.app/ eswar108/rag-architecture-easy-explained-4jpj (дата обращения: 18.08.2024), свободный. – Заглавие с экрана. – Яз. рус.
15. Rag Architecture Easy Explained // Dev. – Режим доступа: https://dev.to/akeshlovescience/rag-architecture-explained-beginner-5hn4 (дата обращения: 06.07.2024), свободный. – Заглавие с экрана. – Яз. рус.
16. Архитектура современных приложений на основе LLM // Хабр. – Режим доступа: https://habr.com/ru/articles/777248/ (дата обращения: 05.09.2024), свободный. – Заглавие с экрана. – Яз. рус.
17. One Hot Encoding in Machine Learning // Geeks for geeks. – Режим доступа: https://www.geeksforgeeks.org/ml-one-hot-encoding/ (дата обращения: 24.05.2024), свободный. – Заглавие с экрана. – Яз. рус.
18. Data Science in 5 Minutes: What is One Hot Encoding? Fawcett A // Educative. – Режим доступа: https://www.educative.io/blog/one-hot-encoding (дата обращения: 07.08.2024), свободный. – Заглавие с экрана. – Яз. рус.
19. Что такое векторная база данных и при чем здесь ИИ // Школа Больших Данных. – Режим доступа: https://bigdataschool.ru/blog/ai-andvector-databases.html (дата обращения: 16.06.2024), свободный. – Заглавие с экрана. – Яз. рус.
20. Как работают Векторные базы данных и Поиск похожих текстов в них // Хабр. – Режим доступа: https://habr.com/ru/articles/784158/ (дата обращения: 23.06.2024), свободный. – Заглавие с экрана. – Яз. рус.
21. Сравнение времени выполнения алгоритма на CPU и GPU // Хабр. – Режим доступа: https://habr.com/ru/articles/525892/ (дата обращения: 24.05.2024), свободный. – Заглавие с экрана. – Яз. рус.
22. NVIDIA Documentation Hub // NVIDIA. Docs Hub. – Режим доступа: https://docs.nvidia.com/ (дата обращения 16.09.2024), свободный. – Заглавие с экрана. – Яз. рус.
23. Знакомство с программно-аппаратной архитектурой CUDA // Библиотека программиста. – Режим доступа: https://proglib.io/p/cuda (дата обращения 09.06.2024), свободный. – Заглавие с экрана. – Яз. рус.

Ссылка для цитирования:

Панкрашов А. С., Окладникова С. В. Интеллектуализация процесса онбординга в строительных компаниях // Инженерно-строительный вестник Прикаспия : научно-технический журнал / Астраханский государственный архитектурно-строительный университет. Астрахань : ГБОУ АО ВО «АГАСУ», 2024. № 4 (50). С. 97–101.