УДК 004.8

РАЗРАБОТКА ОМНИКАНАЛЬНОЙ МУЛЬТИАГЕНТНОЙ ПЛАТФОРМЫ АВТОМАТИЗАЦИИ БИЗНЕС-КОММУНИКАЦИЙ НА ОСНОВЕ ЦИФРОВЫХ ДВОЙНИКОВ КОНТРАГЕНТОВ С МОБИЛЬНЫМ И ВЕБ-ИНТЕРФЕЙСОМ

Хисаметдинова Д.Н. (Университет ИТМО),
Научный руководитель – кандидат технических наук Федоров Д. А.
(Университет ИТМО)

Введение. Современный бизнес сталкивается с растущим объемом коммуникаций через множество каналов (Telegram, VK, email), что требует автоматизации для эффективной квалификации лидов и управления сделками. Применение больших языковых моделей (LLM) с технологией Retrieval-Augmented Generation (RAG) [5] и мультиагентной архитектурой позволяет создать интеллектуальную CRM-платформу, где AI-агенты автоматически ведут диалоги, выполняют бизнес-операции через tool calling и формируют цифровые двойники контрагентов на основе анализа коммуникаций [1-2] и активности в социальных сетях [3]. Однако реализация требует решения архитектурных задач: обеспечения надежности асинхронной обработки критических событий, горизонтального масштабирования при росте нагрузки и интеграции с внешними платформами, а также написания алгоритмов перевода разных представлений данных в оптимальный для языковой модели формат, цепочки вызовов инструментов, а также предоставление всего функционала пользователю в виде интуитивно понятных кроссплатформенного мобильного приложения, веб-сайта и Telegram-бота.

Основная часть. Предлагается распределенная микросервисная платформа, обеспечивающая:

1. Мультиагентную архитектуру с распределением ролей: Intake-agent (прием обращений), Scoring-agent (квалификация лидов), Nurturing-agent (персонализированное сопровождение), RAG-agent (контекстные ответы) и Supervisor-agent (оркестрация) [1].
2. Гарантию доставки критических событий через паттерн Transactional Outbox: атомарное сохранение данных и событий в БД с последующей публикацией в RabbitMQ, что предотвращает потерю сообщений при временных сбоях инфраструктуры [4].
3. Гибридный RAG-подход (BM25 + векторные эмбеддинги + soft-matching) с инкрементальной индексацией в Pinecone, обеспечивающий высокую релевантность ответов агента и постоянную актуальность базы знаний [5].
4. Формирование цифровых двойников контрагентов путем агрегации данных из переписки, соцсетей и веб-активности в векторный профиль, который автоматически встраивается в system prompt LLM для контекстуальной генерации ответов [6-8].
5. Tool Calling для бизнес-операций: автоматическое сохранение информации о клиенте, создание задач в Notion/Asana, добавление событий в Google Calendar, эскалация к человеку через Function Calling OpenAI API [2].
6. Кроссплатформенное мобильное приложение на Kotlin Multiplatform [9], Offline-First подходом (кэширование в SQLDelight с TTL 1 час, автосохранение черновиков каждые 5 секунд), WebSocket real-time обновлениями [10] и push-уведомлениями через Firebase Cloud Messaging [11].

Архитектура использует событийно-ориентированный подход с асинхронной обработкой через RabbitMQ, API Gateway для маршрутизации и полиглот-персистенс

(PostgreSQL, Redis, Pinecone, MinIO) для оптимального хранения данных [4].

Выводы. Спроектированная система реализует парадигму "AI как когнитивная инфраструктура" в отличие от существующих CRM, работающих в режиме "AI как помощник". Мультиагентная архитектура с распределением ролей между специализированными агентами обеспечивает автономную обработку полного цикла B2B продаж — от первого контакта с потенциальным клиентом до закрытия сделки. Концепция агентов с повышенной автономностью впервые применена в CRM-контексте: система агрегирует данные из множества источников (переписка, активность в социальных сетях, веб-поведение) в единый векторный профиль контрагента, который автоматически встраивается в контекст LLM-агента, позволяя генерировать персонализированные ответы с учетом полной истории взаимодействий и бизнес-контекста. Гибридный RAG-подход с инкрементальной индексацией обеспечивает высокую релевантность ответов при постоянной актуальности базы знаний. Система формирует когнитивную экосистему, где каждый агент действует как автономный консультант, менеджер и аналитик, непрерывно обучаясь на собственных взаимодействиях и оптимизируя воронку продаж, что обеспечивает рост конверсии при минимальном человеческом участии.

Список использованных источников:

1. Zhang, S. et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. arXiv preprint arXiv:2308.08155. URL: https://arxiv.org/abs/2308.08155 (дата обращения: 17.12.2025).
2. Brown, T. et al. Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165. URL: https://arxiv.org/abs/2005.14165 (дата обращения: 17.12.2025).
3. Park, J. et al. Generative Agents: Interactive Simulacra of Human Behavior. arXiv preprint arXiv:2304.03442. URL: https://arxiv.org/abs/2304.03442 (дата обращения: 17.12.2025).
4. Richardson, C. Microservices Patterns: With examples in Java. Manning Publications, 2018. — 520 p. — ISBN 978-1617294549.
5. Tao, F. & Zhang, M. Digital Twin Driven Smart Manufacturing // IEEE Access, 2019. Vol. 7. P. 3935-3946. URL: https://ieeexplore.ieee.org/document/8708199 (дата обращения: 17.12.2025).
6. Grieves, M. Virtually Intelligent Product Systems: Digital and Physical Twins // Complex Systems Engineering: Theory and Practice. Progress in Astronautics and Aeronautics, 2019. P. 175-200.
7. Cai, Y., Zheng, Y. et al. Digital Twin-Driven Human–Cyber–Physical Systems // IEEE Internet of Things Journal, 2023. Vol. 10. No. 4. P. 2895-2908. URL: https://arxiv.org/abs/2302.05671 (дата обращения: 17.12.2025).
8. Kotlin Multiplatform Documentation. [Электронный ресурс]. JetBrains. URL: https://kotlinlang.org/docs/multiplatform.html (дата обращения: 17.12.2025).
9. Fette, I. & Melnikov, A. The WebSocket Protocol. RFC 6455. Internet Engineering Task Force (IETF), 2011. URL: https://datatracker.ietf.org/doc/html/rfc6455 (дата обращения: 17.12.2025).
10. Firebase Cloud Messaging Documentation. [Электронный ресурс]. Google. URL: https://firebase.google.com/docs/cloud-messaging (дата обращения: 17.12.2025).