УДК 004.8:336.71

РАЗРАБОТКА ИНТЕЛЛЕКТУАЛЬНОГО АССИСТЕНТА ДЛЯ БАНКА НА ОСНОВЕ ЛОКАЛЬНЫХ МОДЕЛЕЙ ГЕНЕРАТИВНОГО ИИ

Н. С. МОНТИК, Д. А. ИГНАТЮК
Брестский государственный технический университет
Брест, Беларусь

В данной работе рассматривается разработка интеллектуального ассистента для банковского обслуживания, предназначенного для автоматизации консультирования клиентов по банковским продуктам. Описываются архитектурные и технологические решения на основе подхода Retrieval-Augmented Generation (RAG), реализованный прототип системы, а также обосновывается практическая значимость работы. Особое внимание уделяется обеспечению информационной безопасности за счёт использования локальных моделей генеративного ИИ.

Цель работы – разработка прототипа интеллектуального ассистента для банковского обслуживания, обеспечивающего консультирование клиентов по вкладам, кредитам и картам без передачи данных в интернет. Система направлена на снижение нагрузки на контакт-центр и повышение скорости обработки типовых запросов.

Актуальность темы обусловлена необходимостью автоматизации клиентского обслуживания в банковской сфере при жёстких требованиях к информационной безопасности. Существующие облачные ИИ-решения (Gemini, DeepSeek, Claude) не могут быть использованы из-за риска утечки персональных данных. Альтернатива – локальные генеративные модели, работающие в закрытом контуре организации [1].

Существующие банковские ассистенты, как правило, работают по заранее заданным сценариям и не способны генерировать развёрнутые ответы на основе анализа документов. Использование подхода RAG позволяет преодолеть это ограничение за счёт объединения поиска информации и генерации ответа.

Программное средство реализовано в виде веб-приложения, построенного на основе клиент-серверной архитектуры. Архитектурное разделение системы на уровни представления, бизнес-логики и хранения данных обеспечивает модульность, упрощает сопровождение программного кода и позволяет независимо развивать отдельные компоненты системы.

Для реализации выбран подход RAG, объединяющий поиск информации в базе знаний и генерацию ответа на основе найденного контекста [1]. Система включает следующие компоненты:

– векторная база данных FAISS для семантического поиска документов;
– модель эмбеддингов bge-m3 для представления текстов на русском языке [2];
– языковая модель Qwen3.8-Flash-Next для генерации ответов [3];
– фреймворк Ollama для локального запуска моделей.

Входными данными для системы являются текстовые запросы клиентов на естественном языке. На выходе система предоставляет сгенерированный ответ на русском языке, содержащий конкретные условия и рекомендации.

Выбор языка Python обусловлен наличием развитой экосистемы библиотек для работы с ИИ (LangChain, FAISS). В дальнейшем возможна реализация на языке программирования, используемом в инфраструктуре банка.

База знаний формируется из публичных документов банка: условий вкладов, кредитов, карт и тарифов. Документы разбиваются на фрагменты и преобразуются в векторные представления с помощью модели bge-m3. Полученные векторы сохраняются в индексе FAISS, что позволяет выполнять семантический поиск – не по точному совпадению слов, а по смыслу запроса. При обработке запроса текст клиента также преобразуется в вектор, после чего находятся наиболее близкие по смыслу фрагменты документов. Они передаются в языковую модель вместе с системным промптом, задающим правила поведения ассистента. Модель формирует ответ, опираясь на предоставленный контекст, что снижает риск генерации недостоверной информации.

Разработанный прототип демонстрирует способность обрабатывать типовые клиентские запросы в рамках консультационного сценария. Архитектура системы позволяет заменять языковую модель на более тяжёлую без изменения программного кода, что открывает перспективы масштабирования при наличии вычислительных ресурсов.

На текущем этапе прототип имеет ряд ограничений: база знаний охватывает только публичные документы и не интегрирована с внутренними системами. Перспективы развития связаны с расширением базы знаний, интеграцией с CRM-системами банка и добавлением механизмов логирования диалогов.

В дальнейшем планируется проведение пилотного тестирования с участием сотрудников банка для оценки качества ответов и выявления нетипичных сценариев. Полученные результаты позволят уточнить требования к системе и определить направления её доработки перед промышленным внедрением.

Практическая значимость результатов состоит в возможности применения разработанного средства в деятельности банков для повышения эффективности клиентского обслуживания при сохранении информационной безопасности.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks / P. Lewis, E. Perez, A. Piktus [et al.] // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 9459–9474.

2. BGE-M3: Multi-lingual, multi-functionality, multi-granularity text embeddings. – URL: https://bge-model.com/bge/bge_m3.html (date of access: 15.08.2026).

3. On the Design of Qwen3.8-Next Architecture: Evaluation, Efficiency, and Training Stability. – URL: https://arxiv.org/html/2608.30320 (date of access: 15.08.2026).