Агенты для выживания

Уже к 2028 году прогнозируемая доля рутинных процессов под управлением ИИ-агентов может достичь 80%

Автономные ИИ-агенты являются не только технологией самого высокого уровня, способной принимать решения и достигать цели. Они стали необходимым условием технологичного развития. Однако, несмотря на все плюсы ИИ-агентов, у некоторых игроков остаются сомнения в целесообразности их использования. Разбираемся в условиях, сложностях и перспективах применения ИИ-агентов

Фото: Абсолют Банк

Текст
АНТОН ПАВЛОВ,
ЗАМЕСТИТЕЛЬ ПРЕДСЕДАТЕЛЯ ПРАВЛЕНИЯ АБСОЛЮТ БАНКА

Розничный банк — инкубатор для ИИ-агентов

Современный розничный банк функционирует в условиях беспрецедентного давления: регуляторные требования ужесточаются, киберугрозы становятся все более изощренными, а клиенты ожидают бесшовного и персонализированного обслуживания. В таких обстоятельствах традиционные подходы к автоматизации — разовые скрипты, изолированные системы правил и ручные процессы контроля — не позволяют справиться с современной конкуренцией на рынке. Решением может стать агентная архитектура искусственного интеллекта. ИИ-агент — это не просто алгоритм, выполняющий заданные инструкции, это автономная система, способная воспринимать контекст, принимать решения, обучаться на обратной связи и адаптироваться к изменениям без постоянного вмешательства человека. В банковской среде, где каждая транзакция, каждая кредитная заявка и каждый инцидент безопасности требует оперативной реакции, такая автономность является важнейшим преимуществом.

Розничный банк представляет собой идеальную среду для внедрения агентного ИИ по нескольким причинам. Во-первых, банк генерирует колоссальные объемы данных: транзакции, кредитные истории, поведенческие паттерны клиентов. Во-вторых, бизнес-процессы в розничном банкинге хорошо формализованы и поддаются алгоритмизации. В-третьих, регуляторная база (прежде всего требования Банка России, ФСТЭК и ФСБ) задает четкие рамки, внутри которых можно проектировать безопасные агентные системы. Наконец, масштаб розничного банка — миллионы клиентов, тысячи сотрудников, сотни филиалов — создает критическую массу для обучения и валидации агентных моделей.

Ахиллесовы пятки ИИ-систем

Тем не менее в погоне за автоматизацией нельзя забывать о рисках. Внедрение ИИ-агентов в критически важные банковские процессы сопровождается специфическим профилем угроз. Прежде всего, это модельные риски, то есть риски возникновения убытков в результате использования недостаточно точных моделей для принятия решений. Они могут возникнуть в результате систематических ошибок из-за смещения обучающих ИИ данных, концептуального дрейфа или состязательных атак на входные данные. Модельные риски представляют особую опасность по причине их скрытого характера. Но это не все угрозы, с которыми может столкнуться банк при использовании агентного ИИ. Атаки на цепочки поставок ИИ, включая компрометацию предобученных моделей и отравление обучающих датасетов на этапе fine-tuning, формируют новый класс уязвимостей. Для противодействия им требуются специализированные инструменты верификации.

Регуляторы задают проектные требования

Проблема прозрачности автоматизированных решений приобретает критическое значение в банковском контексте. Использование нейросетевых моделей типа «черного ящика» в процессах кредитного скоринга или оценки мошеннических транзакций создает регуляторные риски: клиент может потребовать обоснования отказа, и банк обязан его предоставить.

Регуляторные требования российского рынка задают жесткие рамки допустимых архитектурных решений. Требования Банка России в части управления операционными рисками регламентируются Положением 716-П, информационная безопасность финансовых организаций — стандартами ГОСТ Р 57580, директивы ФСТЭК по защите критической информационной инфраструктуры и ограничения ФСБ на применение криптографических средств иностранного производства совокупно определяют пространство допустимых технологических решений. Защиту персональных данных обеспечивает Федеральный закон № 152-ФЗ. Данные ограничения можно рассматривать как препятствие, а можно — как проектные требования.

Безопасный конвейер

Рассмотрим конкретный пример того, как может быть устроена архитектура безопасного конвейера с учетом существующих рисков и регуляторных требований при использовании современных технологий и решений.

Популярная в настоящее время концепция «горизонтали безопасности» предполагает сквозное встраивание защитных механизмов на каждом уровне стека ИИ-системы — от инфраструктуры хранения данных до пользовательских интерфейсов. Исходя из этого принципа, определим ключевые технологии, которые потребуется предусмотреть в рамках конвейера.

Первая — инфраструктура открытых ключей (PKI), которая обеспечит аутентификацию компонентов системы и подписание артефактов на протяжении всего цикла модели.

Вторая — интеграция с SIEM-системами — будет коррелировать события безопасности из разнородных источников и выявлять аномалии поведения агентов в режиме реального времени. Таким образом обеспечивается непрерывный операционный контроль.

Третья — доверенные среды исполнения (Trusted Execution Environments, TEE), которые сыграют ключевую роль в защите чувствительных вычислений: инференса моделей на персональных данных клиентов и обработки конфиденциальных транзакционных паттернов.

Четвертая — технология Intel SGX или ее отечественные аналоги — позволит создавать изолированные анклавы, в которых выполняются критические вычисления, недоступные для наблюдения извне, в том числе со стороны привилегированного программного обеспечения операционной системы или гипервизора. Это решит принципиальную проблему доверия к облачной инфраструктуре.

Пятая — контроль целостности кода через подписанные артефакты в конвейере CI/CD — представляет

собой обязательный элемент архитектуры безопасного развертывания.

Кроме того, каждый компонент системы должен иметь верифицируемую криптографическую подпись. Система управления версиями ML-моделей (ML Model Registry) с поддержкой неизменяемого журнала аудита (immutable audit log) гарантирует воспроизводимость экспериментов и возможность отката к предыдущим безопасным версиям в случае обнаружения инцидента.

ИИ-агенты в операционном бою

Классификация интентов. При работе ИИ-агентов на этапе входящих запросов — транзакционные инструкции, клиентские обращения, внутренние операционные команды — происходят категоризация намерения и маршрутизация их к специализированным агентам-обработчикам. Многоуровневая система интент-классификации позволяет выявлять паттерны социальной инженерии и мошеннических запросов еще на этапе входного контроля.

NLP и OCR мониторинг. Интеграция технологий обработки естественного языка (NLP) и оптического распознавания символов (OCR) обеспечивает непрерывный мониторинг неструктурированных данных в бизнес-процессах банка. Агенты анализируют в режиме реального времени клиентские договоры, платежные поручения, корреспонденцию и внутренние документы, автоматически выявляют отклонения от стандартных паттернов и потенциальные индикаторы мошенничества или нарушений комплаенса.

Динамическое управление ликвидностью. Динамическое перераспределение ликвидности служит наглядной иллюстрацией агентной работы в действии. ИИ-агент непрерывно отслеживает состояние корреспондентских счетов, прогнозирует потоки платежей на горизонте нескольких часов и автономно инициирует переводы между счетами для поддержания оптимального уровня ликвидности. Скорость реакции агента — секунды против часов при ручном управлении — критически важна в условиях высоковолатильных рыночных сессий.

Контур розничного банка. Практическая интеграция ИИ-агентов в операционный контур розничного банка разворачивается по нескольким ключевым направлениям. Кредитный скоринг является наиболее зрелым направлением применения: агентные системы анализируют не только классические финансовые показатели заемщика, но и поведенческие паттерны (историю транзакций, динамику остатков, частоту обращений в контакт-центр), формируя многомерный профиль кредитного риска с точностью, недостижимой для традиционных скоринговых карт.

> **Розничный банк, последовательно реализующий стратегию агентного ИИ с приоритетом безопасности, трансформируется из финансового посредника в интеллектуальную инфраструктурную платформу цифровой экономики**

Фото: Абсолют Банк

Транзакционный анализ. Транзакционный анализ представляет собой второй критический сценарий применения. Агенты мониторинга в реальном времени анализируют поток платежных операций, выявляя паттерны, характерные для отмывания денег, финансирования терроризма и операционного мошенничества. Многослойная архитектура детекции — от правил на основе пороговых значений до графовых нейронных сетей, анализирующих связи между участниками транзакций, — обеспечивает высокую точность при приемлемом уровне ложных срабатываний.

Барьеры зрелости

Российский банковский сектор демонстрирует существенную неоднородность в уровне готовности к агентному ИИ. Крупнейшие банки — Сбербанк, ВТБ, Альфа-Банк — уже перешли в категорию «пионеров», формируя собственные ML-платформы и центры компетенций. Банки второго эшелона находятся в переходной зоне «последователей»: они внедряют готовые решения вендоров, не имея ресурсов для глубокой кастомизации. Региональные банки в большинстве своем остаются на стадии экспери-

ментальных «пилотов», ограниченных отдельными сценариями использования.

Однако вопрос зрелости в определенной степени сопряжен в том числе с проблемами оценки эффективности технологии. Ключевая проблема состоит в том, что наибольшая часть создаваемой ценности носит превентивный характер: предотвращенные убытки от фрода, ненарушенные регуляторные требования, предотвращенные операционные инциденты. Традиционные финансовые модели с трудом поддаются количественному измерению этих «событий, которых не произошло».

Однако существует специализированная система ключевых показателей эффективности для агентного ИИ, которая включает в себя три кластера метрик. Метрики скорости фиксируют задержку принятия решений в миллисекундах, долю транзакций, прошедших полностью автоматическую обработку (STP-rate), и время от обнаружения аномалии до генерации алерта. Метрики качества оценивают долю ложных срабатываний, индекс дрейфа модели (отклонение текущих предсказаний от baseline) и интегральный показатель объяснимости решений. Третий кластер — метрики принятия — измеряет степень доверия сотрудников к рекомендациям агентов и частоту ручных переопределений.

Оценка зрелости культуры ИИ среди сотрудников банка выходит за рамки технических показателей. Исследования демонстрируют корреляцию между уровнем доверия сотрудников к ИИ-рекомендациям и реальной эффективностью систем: низкое доверие приводит к систематическому переопределению агентных решений, что нивелирует операционные преимущества автоматизации. Программы управления изменениями, прозрачность алгоритмов и механизмы обратной связи становятся критически важными факторами успеха внедрения.

Управление рисками: от теории — к практике

Управление рисками ИИ-агентных систем требует совершенно нового подхода: перехода от статических политик безопасности к динамическим моделям контроля, адаптирующимся к изменяющемуся поведению агентов и операционным условиям.

Моделирование нежелательных действий. Систематическое построение сценариев отклоняющегося поведения агентов охватывает как технические сбои (дрейф модели, ошибки в данных, инфраструктурные аномалии), так и намеренные атаки (adversarial examples, prompt injection, data poisoning). Для каждого сценария определяются вероятность возникновения, потенциальный ущерб и контрольные механизмы. Результатом является тепловая карта рисков агентной системы, обновляемая в реальном времени.

Многоуровневый контроль и аудит. Архитектура контроля реализует принцип «многоуровневой защиты»: каждый агент оснащен встроенными ограничителями действий (guardrails), внешним надзорным агентом-аудитором и централизованным оркестратором, контролирующим соответствие действий агентов политикам банка. Журнал аудита формируется в режиме реального времени с криптографической подписью каждой записи, обеспечивая неопровержимость доказательной базы для регуляторных проверок.

Поэтапное внедрение как стратегия минимизации рисков. Стратегия canary deployment адаптирована для банковских агентных систем: новые версии агентов сначала обрабатывают небольшую долю трафика (1–5%) в режиме «теневого» выполнения — действия агента логируются, но не исполняются. Это позволяет накопить статистику реального поведения перед полноценным вводом в эксплуатацию. Автоматические триггеры rollback обеспечивают немедленный возврат к предыдущей версии при пересечении пороговых значений ключевых метрик безопасности.

Банки, внедрившие многоуровневую систему контроля агентных решений, фиксируют сокращение среднего времени обнаружения инцидентов (MTTD) на 73% и снижение операционных потерь от автоматизированных ошибок на 58% по сравнению с базовым периодом.

Безопасность как конкурентное преимущество

Формула успешной трансформации розничного банка: защищенная агентная архитектура плюс масштабируемая автоматизация процессов равняется устойчивому конкурентному преимуществу. Каждый из компонентов формулы является необходимым, но недостаточным: архитектура без масштабирования остается лабораторным проектом, автоматизация без защиты создает экзистенциальные риски, а оба компонента без операционной зрелости не трансформируются в экономическую ценность.

Переход от экспериментальных проектов к полноценному производственному развертыванию требует зрелой MLOps-практики, выстроенных процессов управления изменениями и четкой регуляторной стратегии. Банки, застрявшие в «пилотном чистилище», рискуют упустить окно технологического лидерства.

При этом важно также переосмыслить безопасность. Необходимо воспринимать ее как источник конкурентного преимущества, а не как центр затрат. Банки с репутацией надежных хранителей данных получают преимущество в привлечении клиентов и выстраивании партнерств.

Наконец, формирование банковской экосистемы, устойчивой к вызовам цифровой эпохи, предполагает выход за рамки отдельного института: участие в отраслевых консорциумах по безопасности, вклад в формирование регуляторных стандартов и открытое взаимодействие с академическим сообществом становятся стратегическими императивами.

Розничный банк, последовательно реализующий стратегию агентного ИИ с приоритетом безопасности, трансформируется из финансового посредника в интеллектуальную инфраструктурную платформу цифровой экономики. Это не технологический выбор — это стратегический императив выживания в условиях неизбежной автоматизации финансовых услуг. Б.О