

Рецепт свиных крылышек

На международном форуме Kazan Digital Week 2024 обсудили взаимосвязь ИИ, Open API, Data Governance и регуляторных активностей Банка России



Текст
ВАДИМ ФЕРЕНЦ,
ОБОЗРЕВАТЕЛЬ «Б.О.»

Панельная дискуссия «Современные технологии, методы прогнозирования и анализа в бизнесе», состоявшаяся 10 сентября 2024 года, доказала как по темам докладов спикеров, так и по содержанию вопросов к ним от делегатов из зала, что с технологической точки зрения многие инновации уже вышли на «плато продуктивности» по методике аналитиков Gartner.

Но это вовсе не означает, что с точки зрения бизнеса все так безоблачно и банкиры готовы в промышленном масштабе внедрять ИИ-модели везде, где только можно. Во многом не решены проблемы внешнего доступа к данным, их совместному анализу и обработке в целях составления прогнозных моделей. Множество вопросов сохраняется к процессам практического перехода к Open Data и внедрению Open API.

Три ценности и 10 трендов

В итоге главной ценностью панельной дискуссии стали три момента. Во-первых, проблемы с данными и ИИ характерны не только для финансового сектора. Представители иных отраслей экономики также сталкиваются с ними, поэтому виден вектор на межотраслевое взаимодействие при деятельном участии регуляторов. Во-вторых, ИИ перестает быть вещью в себе. Напротив, как в свое время персональные компьютеры стали обыденностью на рабочих столах банковских служащих, так

Фото: Ассоциация ФинТех



Участники дискуссии (слева направо): Сергей Голицын (Т1), Максим Травин (Росбанк), Сергей Юдин («Яндекс»), Марианна Данилина (АФТ), Дмитрий Марков (МТС), Михаил Комаров (Ростелеком)

и результаты работы моделей становятся гораздо доступнее за счет массового применения промпт-инжиниринга и составления контекста для грамотных запросов для ИИ. В-третьих, проблемы с ИБ перешли от этапа осознания соответствующих рисков к стадии их купирования и управления ими.

Общее, что объединяет всех: ИИ — это лишь одна из технологий, помогающая достичь определенных целей. И рассматривать экономический эффект от ее внедрения можно лишь в связке с Data Governance, инфраструктурой обмена данными (Open API, Open Data, криптоанклавы), ИБ и учетом особенностей внутренней архитектуры больших языковых моделей (LLM).

Понятно, что сделанные выводы являются «средней температурой по больнице», поскольку у крупных банков своя повестка в области ИИ, у средних — несколько иная, а у малых банков пока есть только один выход — следить за достижениями своих «старших коллег», используя наработанные ими лучшие практики и рекомендации регулятора.

Вероятно, именно поэтому модератором дискуссии выступила руководитель управления стратегии, исследований и аналитики Ассоциации ФинТех **Марианна Данилина**, которая в качестве задела для диалога представила исследование АФТ «10 трендов по ИИ». Первая тройка трендов включает в себя: демократизацию генеративного ИИ, появление новых моделей и сервисов на основе ИИ, а также развитие мультимодального ИИ.

Росбанк: ИБ, федерализация и мультимодальность

Максим Травин, директор департамента централизованного управления данными Росбанка, начал с ответа на вопрос: что такое тренд на развитие федеративного подхода? По его мнению, это «и про компетенции управления данными, и про повышение их доступности, и про инструменты для работы с ними, а также про анализ внутри компании. А еще это подход про управление вычислениями».

Мультимодальность — это объединение в рамках моделей и алгоритмов разных категорий данных, которое требует для своей реализации специфической ИТ-инфраструктуры. Поэтому усложняется процедура сбора и предобработки сырых данных. Кроме того, требуется все больше компетенций, чтобы эти данные соединять друг с другом.

Это становится причиной того, что те банки, которые идут по пути наращивания объема хранимой информации, по сути, развивают внутри себя платформ по управлению данными. Как следствие возникает новый вызов: обеспечение федерализации доступа к этим данным. Ведь все больше компетенций требуется не только для того, чтобы собрать данные, но и для того, чтобы раздать их хотя бы внутренним потребителям, не говоря уже о внешних.

Сейчас активно разрабатываются соответствующие подходы, что означает серьезное увеличение количества команд, занимающихся сбором и обработкой данных с последующим их анализом. Примечателен в этой связи тренд на то, что работа с данными перестает быть исключительно прерогативой ИТ, а переходит на сторону бизнеса. Сейчас от ИТ требуются по большей части платформы и инструменты.

Как бизнес использует данные? Новые кейсы аналитики данных связаны с попытками использования генеративных моделей, а они достаточно сложны для адаптации под нужды пользователей. Первопроходцам всегда сложнее, и не всегда их предложения результативны и ведут к успеху. Отсюда вытекает вывод: отчетливо видны появление элементов федерализации в локальных внедрениях и старт универсальных кейсов, результаты которых можно с успехом переиспользовать в иных направлениях деятельности.

Максим Травин привел пример: «Это создание чат-ботов на основе генеративного ИИ. Мы привыкли, что боты предназначены лишь для общения в разных каналах коммуникаций. На самом деле они могут решать гораздо больший круг задач. Они в состоянии накапливать самую разную информацию, в том числе о документах, соединяя воедино

разные категории данных, а затем с высокой эффективностью помогать организовывать работу с ними. Человеку уже не требуется пропускать через себя огромные массивы информации, чтобы получить простой ответ на какой-либо вопрос. Таким образом, оптимизируются внутренние процессы и сокращается время, необходимое на извлечение смысла из этих данных».

Появляются истории успеха, связанные с автоматическим написанием программного кода. В частности, одно из успешных применений ChatGPT связано с генерацией ПО и повышением качества имеющегося. Это пример универсального кейса, результаты которого становятся доступными для разных команд разработчиков. В итоге масштабирование универсальных кейсов упрощает доступ к накопленной экспертизе, а уровень адаптации к этой технологии неумолимо растет. Каждый последующий кейс дается проще, а также быстрее интегрируется в бизнес-процессы банка.

«Такого рода федерализация от упрощения доступа к данным, их объединения из разных источников, а с другой стороны, до кейсов применения ИИ, несомненно, дает синергию. Это происходит за счет роста компетенций и специализаций как в математике, так и в методах применения ИИ. Кроме того, упрощается доступ к нему со стороны бизнеса, а это повышает его эффективность и удешевляет внедрение. Поэтому нет необходимости инвестировать в среднем 30–40 млрд рублей в год в отдельно взятую технологию с нуля. Это происходит за счет переиспользования того, что уже есть, намного более экономичными способами. Но все это невозможно без обеспечения ИБ и применения норм этики ИИ», — считает представитель Росбанка.

«УРАЛСИБ»: повышение эффективности и open source

Дмитрий Гришин, директор по инновациям банка «УРАЛСИБ», продолжил обсудить тренд развития ИИ, но при этом уточнил: «Все строят экосистемы, а мы занимается core-banking и предоставляем клиентам качественные банковские услуги. Значит, мы говорим исключительно о практических вещах».

По его мнению, если посмотреть на P&L банковских рисков, то их стоимость в кредитовании является одной из двух важных величин (после стоимости фондирования), поэтому банки внедряли системы прогнозирования вероятности наступления дефолта заемщиков. А отраслевая конкуренция в сфере построения актуальных скоринговых моделей привела к появлению в банках систем, базирующихся на ИИ, помогающих оценить стоимость рисков, а также определить лимиты чуть ли не на каждого заемщика. Это стало первым большим применением ИИ в кредитно-финансовой сфере.

Второе и не менее важное направление — предсказание вероятности отклика клиента на предложение. Дмитрий Гришин добавил: «В последние пять — семь лет наметился тренд использования ИИ для повышения эффективности бэк-офисных процессов при общении с клиентами. Например, все каналы цифровых коммуникаций у нас в банке “накрыты” сплошной системой мониторинга. За счет преобразования голоса в текст с последующим его семантическим анализом, оценкой эмоциональной коннотации, проверкой на соответствие скриптам обслуживания удается обеспечить высокое качество обслуживания в дистанционных каналах. В офлайне тоже идет внедрение своих решений этого класса».

Где же сегодня проходит граница применимости ИИ? Позиция «УРАЛСИБА» известна: «Что работает внутри банка, должно там же и оставаться. Поэтому использование внешней облачной инфраструктуры, куда мы могли бы передавать на обработку какую-то клиентскую информацию, безусловно, невозможно. Значит, стоит задача развернуть инфраструктуру ИИ в защищенном контуре банка».

Но разработать с нуля свою собственную систему безумно дорого — это сотни миллионов рублей, огромные вычислительные мощности, большой дата-сет, который необходимо собрать, разметить и т.д. Поскольку главной задачей банка является предоставление финансовых услуг своим клиентам, а не превращение в вендора, за основу была взята open-source-модель с минимальным тюнингом.

При этом решалась и менеджерская задача — найти область имплементации этой модели, чтобы получить от ее работы максимальную пользу: повысить качество клиентского сервиса или эффективность операционных процессов. Самый очевидный пример этой практики — разработка некоего «творческого попкорна», т.е. умение очень быстро создавать различные рекламные креативы, картинки, видео, что сегодня практически ничего не стоит банку, зато моментально приносит результат.

А вот когда возникла необходимость построить большую модель по внутренним бизнес-процессам, во весь рост встала задача убедиться в том, что она стабильно выдает тот же самый результат, который соответствует описанию бизнес-процессов. Увы, не всегда это возможно в силу логики самой LLM-архитектуры. Это первое.

Второй важный момент, который нужно решить при внедрении генеративной модели, это организовать факт-чекинг из-за наличия «галлюцинаций». Необходимо научиться правильно перепроверять полученные результаты, потому что иногда случается иметь дело с удивительно правдоподобными, снабженными цифрами и ссылками, ответами, которые не имеют ничего общего с реальностью.

Третий момент, который важно учитывать при внедрении, это организация ролевой модели доступа. ИБ не бывает много!

Не менее важная задача — снабжение модели «контекстным окном». Для этого необходимо провести эмбединг, т.е. факторизацию накопленных баз знаний, которые существуют в банке, например, в области внутренних нормативных документов или клиентской информации. И этот контекст необходимо передать в LLM вместе с промптом для получения правильного ответа, избавленного от «галлюцинаций» модели.

Поэтому правильно выстроенная процедура работы сотрудников — значительный шаг вперед в плане гуманизации взаимодействия людей с цифровыми решениями: от перфокарт — к интеллектуальным чат-ботам и т.д. Чтобы еще более упростить работу сотрудников банка с ИИ, было решено пойти по пути создания фабрики промптов, которая берет исходный запрос от специалиста, сама пытается понять его смысл, уточняет что-то у че-

ловека в форме наводящих вопросов и далее формирует наиболее эффективный промпт на вход модели, дополняя его контекстом. Затем фабрика проверяет выдачу в соответствии с ролевой моделью на предмет наличия реалистичных ссылок и фактов.

Развитие решения связано с возможностью распространения полученного опыта работы с базой знаний на другие области взаимодействия с клиентом.

Холдинг Т1: от цифровизации к AI-диджитализации

Сергей Голицын, руководитель направления «Т1.ИИ» холдинга Т1, имеющий опыт работы в Сбере и ВТБ, попытался обобщить опыт обоих крупнейших банков с экспертизой IT-компаний: «История мультимодальности ИИ гораздо более сложная, чем обычно принято думать. Если заглянуть в сторону СУБД и КХД, а также инструментов хранения данных Enterprise-уровня, то их всего две или три в мире. При этом эпоха Oracle постепенно начинает уходить в связи с переходом отрасли на микросервисную архитектуру. Понятно и то, что хранить мультимодальные данные в одной СУБД невозможно. Невозможно это и в силу процессов импортозамещения. При этом затрагивается и слой “железа”. Но главное — это умение управлять самими данными».

Любопытно, что Data Governance теперь строится на инструментах ML и ИИ. А сам ИИ начинался с департамента рисков, потом все пошло в CRM, потом — в комплаенс, и сейчас практически нет бизнес-процесса, где не применяются модели. К тому же практически вся работа с данными уходит из бэк-офиса во фронт-офис, поближе к конечным пользователям, меняя фокус бизнеса в целом с цифровизации на AI-диджитализацию.

Какие ключевые барьеры на пути распространении лучших ИИ-практик в целом по финансовой отрасли? По мнению спикера, главный барьер заключается в достаточно слабом уровне применения цифровых платформ, хотя уже появились достаточно мощные отечественные решения, включая MLOps-платформы, помогающие в том числе избежать дефицита «цифровых людей» («D-people» в терминологии Сбера), которые сегодня являются крайне дорогим и дефицитным ресурсом.

По расчетам банкиров, количество этих кадров необходимо в ближайшие годы нарастить в 20–30 раз! Наверное, только MLOps-алгоритмы могут в какой-то мере снизить остроту проблемы (конечно, при условии, что Data Governance в финансовой организации достаточно зрелый).

О каких ИИ-бюджетах можно говорить? Простой пример: при обычном подходе обучения моделей с использованием примерно десятка GPU-карт бюджет может составить 100 млн рублей в год при условии их аренды на какой-либо облачной платформе. Это гораздо дешевле, чем эксклюзивные практики и собственные графические карты.

«Яндекс»: интуиция ИИ и причины «галлюцинаций»

Сергей Юдин, руководитель ML команды «Яндекс Браузера», пояснил, почему все так сложно с точки зрения необходимости промпт-инжиниринга и контекста: «Языковые модели являются вероятностными. Этим они радикально отличаются от классических (детерминированных) моделей, которые работают по четким правилам, формулам и т.д. Машинное обучение входит в это число, поскольку обучение происходит по заранее написанным и изученным правилам. Некоторые скоринговые модели поэтому и строятся на ML, чтобы быть интерпретируемыми».

С появлением генеративных моделей все изменилось: они «рассуждают», основываясь не на фактах, а на некоем неинтерпретируемом человеком представлении о реальности, которое формиру-



Фото: Ассоциация ФинТех

ется при обучении на миллионах текстовых документов. Поэтому сеть пытается сгенерировать максимально правдоподобный, с ее точки зрения, текст. Например, эксперт описал свой собственный опыт общения с LLM от «Яндекса», когда он с удивлением узнал от модели, что существуют рецепты приготовления «свиных крылышек» и даже известна цена этого блюда в ресторанах — 350 рублей за порцию.

На самом деле человек действует почти аналогичным образом, решая какую-то сложную задачу, а это означает, что и у ИИ есть так называемая внутренняя интуиция. Если проанализировать, почему человек внедряет ИИ, то выяснится, что он пытался уйти от решений, подсказанных ему интуицией, к чему-то основанному на четких правилах, например на ML. В случае с LLM-моделями, являющимися одновременно и механизмом логических выводов без четких инструкций, и хранилищем знаний, получается, что человечество пришло к тому, от чего ушло. Но, к счастью, развитие технологического прогресса на этом не остановилось!

Выводы из содержательной беседы

Один из знаковых выводов из прошедшей дискуссии сделал Максим Травин: «Финтех созрел до того, чтобы обсуждать категорию “данные как актив”, которая может нести в себе некую материальную ценность, а также быть способной увеличивать капитализацию компании».

Как следствие эксперты обращают внимание на сдвиг парадигмы: если раньше банкиры собирали данные, структурировали и направляли их в хранилища, то теперь они все чаще нацеливают свою IT-инфраструктуру на применение более сложных алгоритмов,

а также на выстраивание платформ по управлению моделями. Кроме того, все видят, как на финансовом рынке появляются новые подотрасли, которые могли бы переиспользовать накопленный коллегами опыт. Подобные рекомендации для них в скором времени помимо систематизации опыта будут накладывать и определенные ограничения.

Максим Травин утверждает: «Многое из сказанного происходит за кадром для тех, кто смотрит на финтех снаружи. Но внутри, я думаю, в ближайшее время стоит ожидать резкого усиления влияния обозначенных АФТ трендов».

Второй блок выводов связан с Open API и их ролью в управлении данными. По словам Марианны Данилиной, сейчас крупные игроки рынка уже ведут обмен данными между собой в соответствии со стандартами регулятора в рамках экспериментального режима. Планы сформулировать в виде концепции среду Open API в рамках Open Data существуют и, по всей видимости, будут реализованы. В мире это делается по-разному, нам необходимо определиться с нашей моделью, после чего можно ожидать взрыва активности финтеха в этой сфере, как это произошло в свое время в Великобритании.

Сергей Голицын добавил: «Существуют и иные технологии, которые позволяют безопасно объединять данные, например криптографические анклавов, в которые “зашит” такой элемент инфраструктуры ИИ, как AutoML, что помогает строить модели машинного обучения на основе объединенных данных. Эти технологии доступны уже сейчас, их можно применять, к примеру, для разработки банковских продуктов или улучшения клиентского сервиса».

Ссылаясь на недавнюю презентацию **Вадима Кулика**, заместителя президента — председателя правления банка ВТБ, спикер назвал подтвержденный эффект трехлетней цифровой трансформации банка ВТБ — 300 млрд рублей. Кроме того, имеются данные по каждой модели, но пока это закрытая информация.

Модельные подразделения, обслуживающие риски, сегодня имеют не такую высокую долю, как это было недавно. Модельные сервисы заточены на закупки, контакт-центры, работу с нормативными документами. Тем не менее рисквики подсчитали, что повышение на один пункт коэффициента Джини, отражающего качество скоринговой модели, в портфеле кредитов наличными банка ВТБ дает эффект прироста чистой прибыли на 100 млн рублей в год. Так что стимулов развития ИИ более чем достаточно!

БО