Повсеместное внедрение ИИ в бизнес-процессы эксперты называли трендом 2024 года. Массовое внедрение генеративных нейросетей в банках пока впереди, но уже сейчас многие задумываются о практических кейсах использования. О том, какие преимущества и риски могут получить кредитные организации от внедрения больших языковых моделей (LLM), «Б.О» рассказал Александр Крушинский, директор департамента голосовых цифровых технологий компании BSS

Текст
АЛЕКСАНДРА КРЫЛОВА,
ОБОЗРЕВАТЕЛЬ «Б.О»

Александр Крушинский (BSS):

ГенИИ уже может улучшать клиентское обслуживание в банках

Фото: BSS

— Александр, какие новые возможности в обслуживании клиентов открывают перед банками генеративные нейросети и большие языковые модели (LLM)?

— Генеративные нейросети и решения на их основе используются в действующих чат-ботах, голосовых помощниках, в существующих системах речевой аналитики и контент-анализа, и все это они выводят на новый уровень, меняют к лучшему. Так, чат-боты после внедрения в них LLM получают возможность, не прибегая к шаблонам, давать ответ на конкретный вопрос клиента.

Системы речевой аналитики со встроенной большой языковой моделью способны по запросу на естественном языке, без предварительной подготовки шаблона и выделения ключевых слов, проанализировать заданный массив диалогов, найти, к примеру, наиболее частые причины жалоб клиентов и даже выдать рекомендации по улучшению качества обслуживания.

— Чат-боты с LLM обходятся банкам дороже, чем стандартные виртуальные помощники?

— С точки зрения экономики применение генеративных нейросетей в чат-ботах и голосовых помощниках вполне оправдано: создание робота для консультирования клиентов, например, по 100 тематикам даже без интеграции с внутренними системами банка требует кропотливого ручного труда. На подготовку шаблонов для такой нейросети и ее обучение приходится тратить тысячи часов.

А генеративные нейросети умеют читать и создавать текст. Достаточно показать такой сети вопрос клиента и материалы, по которым надо дать ответ, причем в любом формате, и она это сделает. В идеале за день можно создать на основе нейросетей робота, который будет отвечать на любые вопросы клиента, опираясь на весь объем внутренних нормативных документов банка.

— Что же тогда сдерживает проникновение решений на основе генеративных нейросетей в банки?

— Начну с того, что пользоваться самой умной на сегодняшний день большой языковой моделью — ChatGPT от американской компании Open AI — российские банки в силу санкций не могут, а доступные российским корпоративным клиентам генеративные нейросети «Яндекса» и Сбера (YandexGPT и GigaChat) пока не так развиты.

Еще одна проблема, которая удерживает банки от применения этих моделей, заключается в том, что изначально они создавались как облачные. Пока не нужно использовать в ответах на вопросы клиента персональные данные или другую чувствительную информацию, такая модель банк устраивает. Но как только возникает необходимость отвечать предметно и персонализированно, он начинает сомневаться в информационной безопасности использования облачных сервисов. А развертывание больших языковых моделей в контуре обойдется в десятки миллионов рублей для закупки графических карт, необходимых для работы этих моделей, так как они требуют огромного объема вычислительных ресурсов.

Генеративные нейросети умеют читать и создавать текст. Достаточно показать такой сети вопрос клиента и материалы, по которым надо дать ответ, причем в любом формате, и она это сделает. В идеале за день можно создать на основе нейросетей робота, который будет отвечать на любые вопросы клиента, опираясь на весь объем внутренних нормативных документов банка

Другим серьезным фактором, сдерживающим внедрение генеративных нейросетей в обслуживании клиентов в банках, является их склонность к галлюцинированию. Еще три года назад мы не могли себе представить, что генеративные нейросети будут отвечать на вопросы, как человек. Сегодня мы уже знаем, что они «умные», но могут ошибаться. А если нейросеть от лица банка выдает клиенту официальный ответ, который остается в истории его обращений, с ошибкой, это серьезная проблема. Понятно, что со временем проблема будет решена, поскольку есть ряд способов этот риск нивелировать, но пока этого не произошло.

— Вы тоже работаете над этим? BSS же активно развивает свой речевой технологический стек. Тема генеративных нейросетей и LLM вам близка?

— ChatGPT появился примерно осенью 2022 года, и уже в начале весны или лета 2023-го мы поняли, что, хотя эта технология новая и относительно сырая, нужно искать ей применение в банковских процессах. Для начала мы просто добавили в наш чат-бот в одном из банков возможность обращаться к YandexGPT: нажать кнопку и задать любой вопрос этой генеративной нейросети. Мы при этом не передавали ей данные банка и не могли ответить на вопросы, касающиеся его деятельности.

Поскольку доля обращений с попытками использовать LLM для решения бизнес-задач, например создания рекламных текстов, в пуле всех запросов оказалась довольно высока, мы вывели этого бота для клиентов-юрлиц из сегмента малого и среднего бизнеса, а также индивидуальных предпринимателей. И они продолжают им пользоваться.

Следующий шаг — применение генеративной нейросети для ответов на вопросы об услугах и продуктах банка на основе его материалов с использованием технологии RAG (Retrieval Augmented Generation). Мы и сейчас движемся в этом направлении: упаковываем нормативную информацию банка и просим LLM использовать ее в ответе на предметный вопрос. И это серьезное применение, поскольку нейросети задаются те же вопросы, что и оператору-человеку.

Еще один очень важный для ботов с использованием LLM момент — деперсонализация. Если банк применяет нейросеть из облака, то ему необходимо исключить риск утечки персональных данных клиента, а они могут промелькнуть в запросе, даже

Реклама

если мы о них не спрашивали. Клиент может просто позвонить и представиться: «Я Иванов Иван Иванович, пользуюсь вашей картой с таким-то номером, помогите мне решить проблему». Для того чтобы не передавать эти персональные данные в облако, их нужно найти и вычистить перед отправкой.

Цифры из диалога удалить довольно просто, но по-хорошему помимо них нужно захватывать и другие конфиденциальные данные: фамилию, имя, отчество или кодовые слова. А в идеале надо отдавать в нейросеть ответ, очищенный от персональных данных, но с возможностью их восстановления для ее ответа — например, чтобы она могла обратиться к человеку по имени и отчеству.

— Что дает использование технологии RAG?

— Так как генеративная нейросеть умеет формулировать ответы на основе документов, логично было бы отдать ей все нормативные документы банка на сотнях тысяч страниц и при каждом запросе просить изучить их и дать ответ. Но это дорого: чем больше документов подается на вход модели, тем выше стоимость обработки запроса. Кроме того, у генеративных моделей есть ограничение по количеству символов, которое им можно передать. Пока это десятки страниц.

Для того чтобы LLM исходя из содержания документа могла ответить на вопрос клиента, ей необходимо показать его фрагмент, который она должна учитывать. Естественно, для этого нам нужно найти в документе релевантные отрывки. Эту задачу и решает RAG. Он «просматривает» документы, разрезает их на небольшие части, структурирует и укладывает их таким образом, чтобы модели было легче вести поиск с учетом заголовков и перекрестных ссылок. При поступлении запроса от клиента инструмент RAG в нашем сценарном боте с помощью сложных алгоритмов подбирает релевантные запросу фрагменты документов, формирует оптимальный объем материалов и передает его на вход модели для подготовки ответа.

Этот участок до входа в саму генеративную сеть крайне важен как один из способов снизить ее галлюцинирование: чем более точную, компактную и релевантную информацию мы покажем генеративному искусственному интеллекту, тем точнее и правильнее будет его ответ и тем ниже будет уровень галлюцинаций.

— Над чем вы сейчас работаете в части прикладного использования генеративных нейросетей в банках?

— У нас есть один большой проект, и мы находимся на этапе внутреннего тестирования на выделенной группе клиентов. В первом квартале 2025 года планируем вывести генеративную нейросеть на широкую аудиторию, а пока последовательно улучшаем все аспекты обслуживания: работаем над деперсонализацией и улучшением качества ответов, приводим документы банка в определенный вид, чтобы их правильно потом искать, раскладывать.

Мы умеем использовать большую языковую модель для выделения и определения даже тех тематик обращений клиентов в контакт-центр, которые в нее заранее не закладывались. Если какая-то тематика начала часто появляться в запросах, модель сама обратит на нее внимание аналитика и предложит ему разобраться, почему произошел такой всплеск интереса

Занимаемся тюнингом большой языковой модели. Дело в том, что она может вести себя с клиентами несколько панибратски, например переходить в диалоге на «ты». И даже если ей говорить, как надо себя вести, она все равно время от времени «срывается». Такое поведение не подходит для обслуживания enterprise-клиентов, и мы пытаемся с этим бороться путем дообучения самой модели. Показываем ей множество примеров именно вежливого обращения и таким образом стараемся научить ее, чтобы она обращалась к клиентам на «вы», вела себя корректно и вежливо.

В принципе, потенциал для применения этой генеративной нейросети в банках уже очевиден. Процент автоматизации, а это основная метрика, по которой оценивается эффективность робота, показывающая, сколько процентов обращений он готов обработать без перевода на оператора, уже сравним с показателями скриптового бота. При этом для настройки бота со встроенной LLM требуется в несколько раз или даже в десятки раз меньше времени. В целом, наша большая языковая модель умная, хорошо отвечает на вопросы, но нам еще предстоит решить проблемы с деперсонализацией и с минимизацией ответов, вводящих клиента в заблуждение.

— Как вы применяете LLM в речевой аналитике?

— В контакт-центрах кейсов применения больших языковых моделей и генеративных нейросетей побольше. Тут надо пояснить, что LLM обучена на больших массивах данных и потому очень «умная», но она может и не генерировать текст, а решать задачи аналитического характера: классификация, выявление тематик в речи клиента и операторов.

Мы умеем использовать большую языковую модель для выделения и определения даже тех тематик обращений клиентов в контакт-центр, которые в нее заранее не закладывались. Если какая-то тематика начала часто появляться в запросах, модель сама обратит на нее внимание аналитика и предложит ему разобраться, почему произошел такой всплеск интереса. Это первый кейс — кластеризация.

Также мы научились с помощью LLM делать саммаризацию звонков — вычленять суть запроса клиента и действий оператора для его решения и представлять ее в виде одного абзаца. Конечно, все диалоги в контакт-центрах записываются и транскрибируются. Но, как правило, полная расшифровка беседы избыточна, потому что помимо обсуждения целевой проблемы клиента она может содержать довольно много лишней информации. А ее сжатое изложение очень удобно для быстрого анализа диалогов

в контакт-центрах банков. Обращений туда поступает очень много, и их все нужно быстро и эффективно анализировать.

А еще мы можем задавать большой языковой модели вопросы на естественном языке, и она их понимает. Раньше, для того чтобы отобрать звонки, например, во время которых клиенты выражали недовольство, в системе речевой аналитики нужно было настроить маркеры — прописать в них какую-то лексику, которая говорит о том, что заказчик недоволен, а потом с их помощью отфильтровать диалоги. И только тогда система отберет звонки, в которых клиент использовал заданные слова и выражения. Большой языковой модели можно просто написать: отбери и покажи мне все звонки, в которых клиент ругался. Благодаря тому, что LLM в целом приближаются к уровню интеллекта человека, система, в которой они используются, сама выявит и покажет такие диалоги. Можно поставить перед LLM более сложную задачу: попросить найти и показать диалоги, в которых клиент отказался от предложения, к примеру, кредитной карты, а оператор не предложил ему те условия, которые должен был предложить по скрипту. И нейросеть поймет этот запрос, сформулированный на обычном человеческом языке, и решит эту задачу.

Помимо ботов и речевой аналитики у нас есть продукт «База знаний» — wiki-образная система для размещения в контуре банка, в которой он может хранить всю информацию в структурированном виде с гиперссылками, чтобы сотрудники могли ею пользоваться. Использование генеративных моделей в «Базе знаний» позволяет отвечать на вопросы клиентов по статьям и документам, хранящимся в ней, просто и быстро.

Теперь, если нашу «Базу знаний» клиент спросит: «Какой тариф по депозиту?», она ответит ему: «13%, детализацию можно посмотреть по ссылке». Это еще одно крупное применение генеративной нейросети в наших продуктах.

— Бот со встроенной LLM для операторов контакт-центра — помощник или конкурент?

— Конечно, помощник, ведь он может выступать в роли суфлера. Когда оператор обслуживает клиента в текстовом канале, робот-суфлер дает ему подсказки, предлагая уже готовый текстовый ответ, который достаточно подтвердить одним кликом или поправить, если ответ не нравится. Когда оператор обслуживает голосовой вызов, суфлер в текстовом режиме выдает ему релевантные подсказки, которые можно использовать при общении.

Если мы говорим о речевой аналитике в контакт-центре, то в этом сценарии система с большой языковой моделью помогает супервайзеру и играет скорее роль контролера качества работы операторов. Но надо признать, что появление в контакт-центре робота с возможностями генеративной нейросети имеет следствием сокращение числа операторов. Я не думаю, что этого стоит пугаться, тем более что у сотрудников контакт-центров довольно много рутинной и малоприятной работы. Понятно, что именно такие процессы в первую очередь автоматизируются, а операторы, которые в них участвовали, сокращаются. Зато у операторов появится больше времени заниматься сложными вопросами, с которыми не справляется бот. И работа в этом случае становится для них более интеллектуальной, более интересной.

В идеальном мире операторы остаются для решения суперсложных вопросов, которыми интересно заниматься. Это становится квалифицированной, может быть, даже немассовой деятельностью. Робот помогает оператору ответить на простой вопрос, другая нейросеть говорит ему, что ему можно улучшить в работе, чтобы обслуживать клиента более качественно. Третья — дает ему подсказки «в моменте», чтобы ему ничего не приходилось искать руками. И с этой точки зрения нейросеть становится помощником человека.

Помимо ботов и речевой аналитики у нас есть продукт «База знаний» — wiki-образная система для размещения в контуре банка, в которой он может хранить всю информацию в структурированном виде с гиперссылками, чтобы сотрудники могли ею пользоваться. Использование генеративных моделей в «Базе знаний» позволяет отвечать на вопросы клиентов по статьям и документам, хранящимся в ней, просто и быстро

— Видите ли вы запрос банков по внедрению генеративных нейросетей и больших языковых моделей для использования в их клиентском сервисе?

— Интерес к использованию технологий искусственного интеллекта для обслуживания клиентов у банков огромный. У многих уже сформировались ожидания от внедрения генеративных нейросетей, часто завышенные. Сталкиваемся мы и с опасениями. Те, для кого искусственный интеллект — это некий черный ящик, которому страшно верить, еще не готовы предпринимать практические шаги к его внедрению.

Как и в любой другой модели появления новой технологии и нового продукта, среди банков есть новаторы, которые постоянно экспериментируют с чем-то, потому что это интересно, потому что они понимают, что нужно всегда заниматься всем новым, чтобы быть первыми. У нас есть такие клиенты, которые готовы торить эту дорогу вместе с нами, чтобы быть впереди. У новаторов есть ранние последователи, готовые подхватить знамя и внедрять инновационные решения, имея перед собой хотя бы один удачный пример.

В принципе, мы сейчас находимся в этой фазе. Удачные примеры уже есть, и мы довольно активно предлагаем решения банкам, а они активно ими интересуются. Я думаю, довольно быстро мы придем к тому, что генеративные нейросети, в частности большие языковые модели, начнут использовать все. Тем более что это с точки зрения скорости внедрения, с точки зрения затрат — более простая история, чем реализация обычного голосового или чат-бота. Б.О