У меня зазвонил телефон. Кто говорит?

Интерес российских банков к большим языковым моделям, LLM, понятен: американская компания OpenAI объявила, что количество активных пользователей ChatGPT в августе превысило 200 млн, а компания Gartner недавно спрогнозировала, что в 2025 году IT-директора по всему миру начнут инвестировать в генеративный искусственный интеллект. В этих условиях важно не оказаться на позиции догоняющих и попытаться поставить на службу генеративную нейросеть, пусть даже российскую

Текст
АЛЕКСАНДРА КРЫЛОВА,
ОБОЗРЕВАТЕЛЬ **«Б.О»**



В России собственные генеративные нейросети и большие языковые модели, которые очень интересуют компании из самых разных отраслей экономики, в том числе из банковского сектора, есть у «Яндекса» и Сбера. Каковы же сценарии применения LLM, критерии оценки их эффективности, проблемы их использования, которые являются продолжением их достоинств, а также перспективы применения генеративных моделей в банках?

Итак, Large Language Models — это одно из приложений искусственного интеллекта, служащее основой для решения задач компьютерной обработки естественного языка. Благодаря заложенным в них алгоритмам искусственного интеллекта LLM могут «читать» и «понимать» тексты, написанные людьми, а также создавать свои собственные.

Поиск применений больших языковых моделей и проработку их сценариев сейчас ведут многие разработчики и системные интеграторы, заручившиеся поддержкой банков-новаторов. Поскольку все они следуют в одном русле, в их подходах и разработках

немало общего. Неудивительно, что комментарии участников рынка относительно использования LLM банками в клиентском сервисе как будто дополняют друг друга.

«Сейчас мы наблюдаем активное исследование технологии и некоторое количество пилотных проектов, однако полноценных решений на рынке еще мало. Бизнесу еще предстоит найти ключевые ценности от интеграции технологий LLM, в том числе такие, которые окупят их внедрение», — констатировала **Александра Деханова,** руководитель направления диалоговых систем Naumen. По словам эксперта, большинство гипотез о сценариях применения больших языковых моделей в банках сконцентрировано сегодня в клиентском сервисе и улучшении опыта клиентов, а также в оптимизации внутренних процессов. «Другая плоскость применения — quality management, речевая аналитика», — добавила Александра Деханова.

Анна Власова, руководитель отдела компьютерной лингвистики ГК «Наносемантика», объяснила востребованность больших языковых моделей в обслуживании клиентов банков тем, что они идеально подходят для коммуникаций в форме «вопрос — ответ», хотя не только для них. По опыту создания сложных текстовых и голосовых роботов для крупных компаний, в том числе из банковской сферы, накопленному «Наносемантикой», наиболее популярными являются такие диалоговые сценарии использования LLM:

- информация по предложениям и услугам конкретного банка, по работе подразделений и контактной информации;
- общая информация по финансовым вопросам, условиям проведения транзакций, платежным системам;
- детализация вопроса клиента, поиск необходимой для ответа информации в документации банка;
- передача диалога оператору, если решить проблему силами LLM невозможно;
- работа с эмоциональной окраской обращений клиентов.

«LLM может определять агрессию и пытаться ее снизить в процессе диалога. Даже если разговор придется передать оператору, то клиент основную агрессию уже сбросит на виртуального ассистента с LLM, а не донесет до сотрудника банка», — пояснила Анна Власова.

Еще одно применение генеративных моделей — добавление их в такой продукт, как база знаний. В компании Naumen тоже запущен GPT-ассистент для текущих пользователей базы знаний Naumen KMS. «Это умный помощник для помощи операторам, который работает синхронно с поиском и позволяет искать информацию быстрее, получать сгенерированные понятные и точные ответы без глубокого поиска, сокращая время обслуживания клиентов», — пояснила Александра Деханова.

Повышение «интеллектуального уровня» имеющихся в банке диалоговых систем — самое понятное и логичное место применения генеративных нейросетей. С добавлением в эти системы LLM у них появляется способность общаться и создавать контент на естественном языке, что существенным образом упрощает взаимодействие с операторами, клиентами, супервайзерами, аналитиками и другими пользователями.

И хотя пока эта работа носит по большей части экспериментальный характер, а проекты по внедрению генеративных сетей находятся в стадии пилотирования, уже сегодня можно говорить о критериях оценки возможного эффекта появления нейросетей нового поколения в банках.

Сегодня понятно, что чат-боты с LLM на службе в клиентском сервисе способны повысить удовлетворенность клиентов обслуживанием и их лояльность, в системах речевой аналитики с генеративными моделями проще и дешевле разрабатывать и развивать, а «поумневшие» базы знаний становятся инструментом повышения эффективности сотрудников. Но как оценить эти результаты в цифрах?

В компании Naumen предлагают оценивать уровень удовлетворенности клиента (индекс CSI) от взаимодействия с диалоговым роботом. «LLM также потенциально могут влиять на оптимизацию средств за счет сокращения времени обслуживания клиентов (в частности, если технология используется как вспомогательная к работе оператора)», — отметила Александра Деханова. И это вполне можно измерить.

По одному параметру оценить пользу внедрения новой технологии для бизнеса сложно, считает Анна Власова. Вместе с тем комплексные методики оценки эффекта от использования LLM для консультаций клиентов банка уже есть. «В любом случае считается экономия в деньгах и в человеческих ресурсах. Какую долю коммуникаций берет на себя языковая модель, можно прикинуть, например, по количеству переводов на оператора. Чем их меньше, тем больше человеческих ресурсов экономит модель. Эффективность продаж оценивается так же, как эффективность продаж людьми, и можно посчитать, насколько успешно справляется виртуальный ассистент с подключенной LLM», — пояснила эксперт.

Но для того чтобы дело дошло до использования методик оценок эффективности применения больших языковых моделей в обслуживании клиентов, еще предстоит преодолеть несколько барьеров, связанных как с технологическими особенностями самих генеративных нейросетей, так и с их адаптацией к условиям конкретных кредитных организаций.

«LLM — это искусственный интеллект, который в результате обучения в некотором смысле формирует мнение по различным банковским вопросам. Интеллектуальная модель может посчитать более подходящим не тот ответ, который выдал бы человек, или даже сгенерировать совершенно новый текст, который может оказаться неполным или некорректным», — пояснила Анна Власова.

Скорректировать поведение модели можно путем ее дообучения или даже переобучения, но полностью контролировать генерацию ответа невозможно, это не предусматривает сама технология. «Нельзя выписать все ответы, которые может дать LLM, и зафиксировать этот список», — констатировала Анна Власова.

Еще одна проблема заключается в том, что у банков особо строгие требования к безопасности данных, а у больших языковых моделей — к производительности вычислительных ресурсов. И это противоречие не позволяет идти привычным путем — устанавливать LLM в контур кредитных организаций. «Для использования больших языковых моделей в чат-ботах необходима

МНЕНИЕ ЭКСПЕРТА



Александр Крушинский,
директор департамента голосовых цифровых технологий компании BSS

Галлюцинирование генеративных моделей — серьезная проблема, поскольку она свидетельствует об отсутствии управляемости. Сценарные боты, даже неправильно поняв вопрос, будут отвечать строго по шаблону, и мы знаем, что от себя они ничего не скажут, а генеративная нейросеть способна к такой самодеятельности, так как это свойственно ее природе.

Цена ошибки в банковском обслуживании, которая может произойти вследствие галлюцинирования генеративной модели, очень велика. Да, и человек ошибается, но наша задача — сделать так, чтобы генеративный ИИ в этом вопросе стал лучше человека, чтобы генеративные модели можно было эффективно применять с учетом ограничений и требований, которые есть у enterprise-заказчиков из финсектора, в том числе в части проблем с неверными ответами.

В нашей методике тестирования все ответы генеративной нейросети делятся на бóльшее количество категорий, чем просто правильные и неправильные. Есть ответы правильные. Есть ответы правильные, но либо неполные, либо избыточно полные, в которых LLM не соврет, но что-то недоговорит или, наоборот, скажет слишком много. Есть ответы, неадекватные вопросу, но не вводящие в заблуждение. К примеру, спросит человек о процентной ставке по депозиту, а нейросеть ему расскажет о погоде. И самая опасная категория — это ответы, вводящие в заблуждение: когда, например, человек спрашивает, какой процент по депозитам, а нейросеть почему-то отвечает не 20, как на самом деле, а 12. Такая ошибка для клиента неочевидна, он введен в заблуждение и либо не воспользуется услугой и уйдет, либо совершит какое-то невыгодное для него действие. Пусть не все ответы генеративной модели будут полными, но галлюцинирование хотя бы в этой части нужно свести к минимуму.

Сейчас доля ошибочных ответов нашей большой языковой модели по разным категориям тематик «гуляет» в диапазоне от 1 до 8%. У нас цель — сузить диапазон до 0,5–1%, чтобы языковую модель можно было смело использовать в банках.

Нам важно приблизиться к тому, чтобы генеративная нейросеть совершала такие ошибки реже человека. И мы постепенно идем к этой цели, потому что умеем тестировать LLM и тратим на это много времени, а также потому, что контролируем вход модели и проверяем ее выход — смотрим, насколько он соответствует источнику.

достаточно высокая скорость работы для быстрых ответов, что особенно критично в кейсах помощи оператору внутри диалога с клиентом. Банкам требуются собственное "железо" и достаточные вычислительные мощности, GPU, что приводит к значительным, ощутимым даже для крупных банковских структур, расходам при внедрении», — отметила Александра Деханова.

Работы по оптимизации вычислительной инфраструктуры для LLM ведутся. К примеру, в компании BSS адаптировали генеративную модель к работе на центральных процессорах серверов вместо графических. «За два года требования наших моделей к производительности "железа" снизились раза в четыре», — рассказал **Александр Крушинский**, директор департамента голосовых цифровых технологий компании BSS. Работа идет довольно активно, однако, по его словам, генеративный искусственный интеллект по-прежнему потребляет очень много ресурсов.

Альтернативой размещения больших моделей на вычислительных мощностях банка является использование большой языковой модели из облака. Но в этом случае для обеспечения безопасности чувствительных данных из запроса к LLM их нужно вычищать, то есть деперсонализировать его. «Надо найти и убрать реальные фамилии, лицевые счета, номера договоров, карт и так далее. И такую работу не получится передать разработчикам LLM», — пояснила Анна Власова.

То же самое, напомнила она, нужно делать при дообучении модели на документах банка, подготовленных по технологии RAG (Retrieval Augmented Generation). В документах тоже могут содержаться персональные данные, и их тоже нужно модифицировать.

В целом, эксперты имеют схожие представления о том, какие проблемы еще предстоит решить и как, для того чтобы большие языковые модели могли заступить «на службу» в контактные центры банков. Это внушает уверенность в том, что через какое-то время такое внедрение произойдет.

«Если говорить о будущем LLM применительно к обслуживанию клиентов, то в этой сфере появятся роботы, которые не ошибаются, мгновенно отвечают на интересующий клиента вопрос, могут помочь ему получить услугу, например открыть депозит прямо в процессе диалога», — дал прогноз Александр Крушинский. Также он ожидает появления интеллектуальных систем речевой аналитики, способных самостоятельно выделять из общения с клиентом проблемные моменты и подсказывать, каким образом можно улучшить бизнес-процессы».

В то, что большие языковые модели будут использоваться в банках, верит и Анна Власова. Для этого потребуются особые инструменты — платформы. «Такие платформы позволяют создать виртуального ассистента, который будет использовать LLM, но не ограничиваться этим. Такие платформы удобны для подключения дополнительных нейросетевых сервисов: анализа сентимента (эмоций в диалоге), маскировки персональной информации, голосовых сервисов (распознавания и синтеза речи), выделения в диалоге заданных объектов и многого другого», — пояснила она и добавила, что все эти задачи уже сегодня решает один из продуктов «Наносемантики».

«Влияние технологий будет углубляться: LLM позволят автоматизировать более сложные и составные процессы, решать проблемы нехватки кадров, данных и аналитических мощностей, что в перспективе революционным образом изменит подход к обслуживанию клиентов и продаже и продвижению банковских услуг», — добавила Александра Деханова.

Словом, большие языковые модели — это будущее клиентского сервиса в банках, к наступлению которого готовятся компании-разработчики и интеграторы и о котором стоит начать задумываться банкам. Б.О