ДИАЛОГИ В ГЕНЕРАТИВНОМ КОНТЕНТЕ

ОБЛАСТЬ ПРИМЕНЕНИЯ ГЕНЕРАТИВНОГО ИИ И БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ (LLM) РАСТЕТ. ПО ОЦЕНКАМ GRAND VIEW RESEARCH, ОБЪЕМ ЭТОГО РЫНКА ПРЕВЫСИЛ $4,35 МЛРД И БУДЕТ РАСТИ НА 35,9% ЕЖЕГОДНО В ПЕРИОД С 2024 ПО 2030 ГОД. ПЕРЕХОДЯЩИЙ ТРЕНД ИЗ 2024 ГОДА В 2025 ГОД — ИСПОЛЬЗОВАНИЕ LLM И ГЕНЕРАТИВНОЙ НЕЙРОСЕТИ ДЛЯ УЛУЧШЕНИЯ КЛИЕНТСКОГО ОБСЛУЖИВАНИЯ. О ВОЗМОЖНОСТЯХ И РИСКАХ ПРИ ВНЕДРЕНИИ ДАННЫХ ЦИФРОВЫХ ИНСТРУМЕНТОВ РАССКАЗАЛ КОРРЕСПОНДЕНТУ GUIDE ДИРЕКТОР ДЕПАРТАМЕНТА ГОЛОСОВЫХ ЦИФРОВЫХ ТЕХНОЛОГИЙ КОМПАНИИ BSS АЛЕКСАНДР КРУШИНСКИЙ. АЛИСА МАЛЬЦЕВА

GUIDE: Какие задачи бизнеса могут закрыть сегодня генеративный ИИ и LLM?

АЛЕКСАНДР КРУШИНСКИЙ: Сегодня генеративный ИИ и LLM стали хорошим инструментом для практического использования в компаниях, где важным фактором является клиентское обслуживание. Одна из основных функций — автоматизировать рутинные процессы и улучшить качество сервиса. Такие продукты на основе генеративной нейросети уже востребованы в финансовой сфере, контакт-центрах, телекоме, электронной коммерции и некоторых других сегментах. Я думаю, довольно быстро мы придем к тому, что генеративные нейросети начнут использовать все. Тем более что с точки зрения затрат и скорости внедрения — это более простая история, чем реализация обычного голосового или чат-бота.

Генеративные нейросети и решения на их основе выводят чат-боты, голосовых помощников на новый уровень. В частности, роботы способны дать ответ на конкретный заданный вопрос клиента. Ведь генеративные нейросети могут создать текст, обработав массив материалов, которыми зачастую выступают внутренние нормативные документы.

Также робот может рассказать о продуктах и услугах организации, создав легкий рекламный текст. А системы речевой аналитики с LLM со встроенной большой языковой моделью способны проанализировать заданный массив диалогов как текстовых, так и речевых, найти наиболее частые причины жалоб клиентов и выдать рекомендации по улучшению качества обслуживания.

G: Что послужило стимулом развития собственных продуктов?

A. K.: Мы увидели большой потенциал развития технологии ChatGPT. Для начала добавили в наш чат-бот в одном из банков возможность обращаться к YandexGPT. Увидели положительный отклик. Далее, используя технологию RAG (Retrieval Augmented Generation), обучили генеративную нейросеть отвечать на вопросы об услугах и продуктах банка. При поступлении запроса от клиента инструмент RAG в нашем сценарном боте с помощью сложных алгоритмов подбирает релевантные запросу фрагменты среди нормативных документов заказчика, и передает их на вход LLM для подготовки ответа.

При этом, чтобы не допустить передачи персональных данных во внешнюю LLM (в случае если используется облачная LLM, а не модель внутри контура заказчика), данные анонимизируются перед отправкой.

Помимо клиентского бота, мы встроили AI-функционал в нашу «Базу знаний» — wiki-подобную систему, в которой компания может разместить внутренние документы. Теперь сотрудник или клиент вместо того, чтобы искать ответы на свои вопросы в этих документах, может просто задать вопрос «Базе знаний» и получить конкретный ответ со ссылкой на исходные документы, на которых этот ответ основан.

В первом квартале 2025 года планируем вывести генеративную нейросеть на широкую аудиторию. А пока последовательно улучшаем все аспекты обслуживания: обучаем вежливому общению, работаем над деперсонализацией и минимизацией ответов, вводящих клиента в заблуждение.



МАКСИМ ГОРИН

G: Только ли на банковскую сферу будет рассчитан новый проект?

A. K.: Нет, в принципе решения на базе генеративного ИИ могут помогать в любых областях и подходят для любых масштабов. Если говорить конкретно о наших продуктах, то мы ориентируемся на средний и крупный enterprise, поэтому наша бот-платформа и платформа речевой аналитики предназначены для относительно крупных контакт-центров (от нескольких десятков до нескольких тысяч операторов), а «База знаний» рассчитана на организации с тысячами и десятками тысяч сотрудников.

Более того, мы стараемся учитывать отраслевую специфику основных сегментов. Сегодня мы предлагаем отраслевые решения для финсектора, телекома, страховых компаний, крупного ритейла, медицинских и энергосбытовых организаций, а также для региональных линий МФЦ и службы 122.

G: Какие есть риски внедрения генИИ?

A. K.: При возникновении необходимости персонализированного ответа появляется риск информационной безопасности, ведь в таком случае используются персональные данные клиента. Еще один сдерживающий фактор — склонность генеративной нейросети к галлюцинированию, иначе к возможным ошибкам. И если нейросеть от лица, например, финансовой организации, выдаст клиенту официальный ответ на обращение с ошибкой, это может обернуться серьезной проблемой. Безусловно, и человек может допустить ошибку. Но сегодня наша задача сделать так, чтобы генеративный ИИ в этом вопросе стал лучше человека.

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%. Нам важно приблизиться к тому, чтобы генеративная нейросеть совершала такие ошибки реже человека.

G: Как большая языковая модель применяется в деятельности операторов контакт-центров?

A. K.: LLM обучена на больших массивах данных и может решать задачи аналитического характера: классификации, выявления тематик в речи. В частности, эти функции находят применение в контакт-центрах. Также мы научились с помощью LLM делать суммаризацию звонков: вычленять суть запроса клиента и действий оператора для его решения и представлять ее в виде одного абзаца. Такое сжатое изложение диалога удобно для быстрого анализа, ведь обращений, как правило, в контакт-центры поступает много.

Также можно попросить большую языковую модель отобрать звонки, в которых человек выражает недовольство, отказывается или соглашается на условия или ему были предложены не все варианты. Нейросеть поймет запрос, сформулированный на обычном человеческом языке, и решит эту задачу. Таким образом, мы можем отслеживать качество обслуживания.

А совсем скоро мы выпустим новый продукт на базе AI — тренажер оператора, который эмулирует живой диалог с клиентом, позволяя операторам контакт-центра тренировать практические навыки отработки сложных ситуаций «на роботе», чтобы уверенно решать их при общении с живым клиентом.

G: Как в условиях кадрового дефицита может помочь бот со встроенной LLM?

A. K.: Бот со встроенной LLM может автоматизировать рутинные процессы и высвободить человеческий ресурс на другие более сложные задачи. Работа в этом случае становится для операторов более интеллектуальной, более интересной.

В то же время стоит расценивать робота и как помощника оператора. Например, при обслуживании клиента в текстовом канале робот-суфлер дает оператору подсказки, предлагая уже готовый ответ, который можно или подтвердить, или поправить. Когда оператор обслуживает голосовой вызов, суфлер в текстовом режиме готов выдать 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 роль контролера качества работы операторов.

G: Как в будущем может расшириться сфера применения генеративного ИИ и LLM?

A. K.: С одной стороны, с учетом текущего состояния и невероятно высоких темпов развития технологий сфера применения генИИ практически безгранична. С помощью больших генеративных моделей можно автоматизировать не только рутинные операции, но и творческую деятельность, которая ранее казалась неавтоматизируемой. И сейчас мы видим настоящий бум приложений и сервисов на базе AI.

Например, сейчас набирает популярность агентный ИИ, который, в отличие от традиционных приложений (традиционные, по текущим меркам,— это примерно те, которые были популярны всего лишь год назад), может не просто реагировать на команду, а понять цель, построить план ее достижения и постараться выполнить ее. В процессе такие агенты могут задавать уточняющие вопросы, смотреть что-то в интернете, вызывать внешние сервисы и даже обращаться за помощью к другим AI-агентам.

Одновременно с этим внедрение генеративного ИИ в реальные бизнес-приложения идет достаточно осторожно. Эта осторожность вызвана тем, что AI все же допускает ошибки, а для бизнеса нужен высокий и гарантированный уровень качества. Хорошая новость в том, что сами модели становятся ощутимо качественнее (буквально с каждым месяцем), а разработчики продуктов на базе AI (и мы в том числе) учатся исправлять или обходить типовые ошибки AI и обеспечивать необходимое качество. ■