Как искусственный интеллект влияет на финансы предприятий? 1/19/24

Специалист по искусственному интеллекту отдела ИТ-консалтинга, PwC Латвия
Gunda Karnīte

Генеративный искусственный интеллект (GenAI) стал важным инструментом предпринимательской деятельности, помогая предприятиям оптимизировать процессы, повысить эффективность и сократить издержки. Однако, чтобы понять влияние GenAI на финансы в полной мере, важно рассмотреть издержки данного инструмента с разных точек зрения.

Каждый раз, когда модель GenAI получает запрос (prompt), она задействует большие языковые модели, используя для вычислений графические единицы процессоров. Процесс, при котором введенные данные обрабатываются, известен как обработка естественной речи, а данные измеряются в жетонах (tokens). Каждый жетон соответствует приблизительно четырем знакам английского языка, и с помощью 1000 жетонов можно обработать около 750 слов. Для генерирования текста в модели GPT-4 стоимость 1000 жетонов для ввода текста составляет около 0,03 доллара США, а 1000 жетонов для вывода текста – около 0,06 доллара США. Например, если на предприятии 1000 сотрудников, каждый из которых в день отправляет десять запросов объемом 300 слов каждый, генерируя результат в 100 слов, затраты на генерирование текста в день могут составить 198 долларов США, т.е. около 45 000 долларов США в год.

Большую часть издержек образует настройка (fine-tuning) модели GenAI к определенной задаче или домену (области знаний), чтобы модель могла сгенерировать желаемый результат. Например, модель GenAI, обученную проводить финансовый анализ по данным годового отчета, нужно перенастраивать из-за изменений налогообложения. Настройка модели включает в себя как адаптацию параметров модели, так и использование нового набора данных, чтобы сделать выдаваемые моделью результаты более точными в контексте налогов соответствующего государства. Например, OpenAI использует формулу, чтобы подсчитать суммарные издержки на настройку модели.

Суммарные издержки на настройку равны базовой плате за 1000 жетонов, умноженной на количество вводов жетонов в файл и на количество итераций, необходимых для настройки модели. В терминологии AI итерации определяются как эпохи (epochs). Например, если нужно сделать десять итераций с 50 000 жетонов в каждой и цена 1000 жетонов составляет 0,008 доллара США (плата за обучение модели gpt-3.5-turbo), суммарные издержки на настройку составят:

10 * 50 000 / 1000 * 0,008 = 400 USD.

Сооружение инфраструктуры крупных языковых моделей требует существенных вложений, поэтому большинство предприятий предпочитает использовать облачные услуги, чтобы на их основе приводить в действие модели GenAI. Издержки облачных услуг состоят из нескольких позиций: плата за время использования, плата за хранилище, плата за инстанцию общего значения, плата за инстанцию оптимизированного вычисления, цена годового абонемента. Крупнейшими поставщиками публичных облачных услуг на данный момент являются Amazon Web

Services (AWS), Azure, Google Cloud Platform и Oracle, у каждого из которых цены по конкретной позиции издержек отличаются.

Основной аргумент при выборе архитектуры облачных услуг для модели GenAI – будет ли она приводиться в действие на публичном облаке, частном облаке или многооблачных решениях. К примеру, медицинским учреждениям важно обеспечить защиту данных пациентов, поэтому более подходящими на первый взгляд будут частные облачные хранилища, которые вначале могут казаться выгоднее публичных, но в долгосрочной перспективе могут повлечь дополнительные расходы на содержание и обновление моделей.

Общие издержки использования и настройки GenAI могут отличаться в зависимости от многих факторов, включая используемые услуги и спецификации модели. Важно подчеркнуть, что одним из основных факторов в создании моделей GenAI является качество данных, поскольку без высококачественных данных модель GenAI может требовать неоднократной настройки, приводя к повышению издержек и потребления ресурсов. Чтобы уточнить расчеты, необходимо принять во внимание и обучение рабочей силы использованию GenAI, меры по обеспечению безопасности данных и этические соображения. И, конечно, важно всегда консультироваться с отраслевыми экспертами для определения наиболее эффективного и выгодного предприятию использования GenAI.

Чтобы подробнее узнать о технологии GenAI, приглашаем 23 мая текущего года на бесплатный вебинар PwC’s Academy «Практическое применение искусственного интеллекта (GenAI)».