Большая языковая модель (Large Language Model)

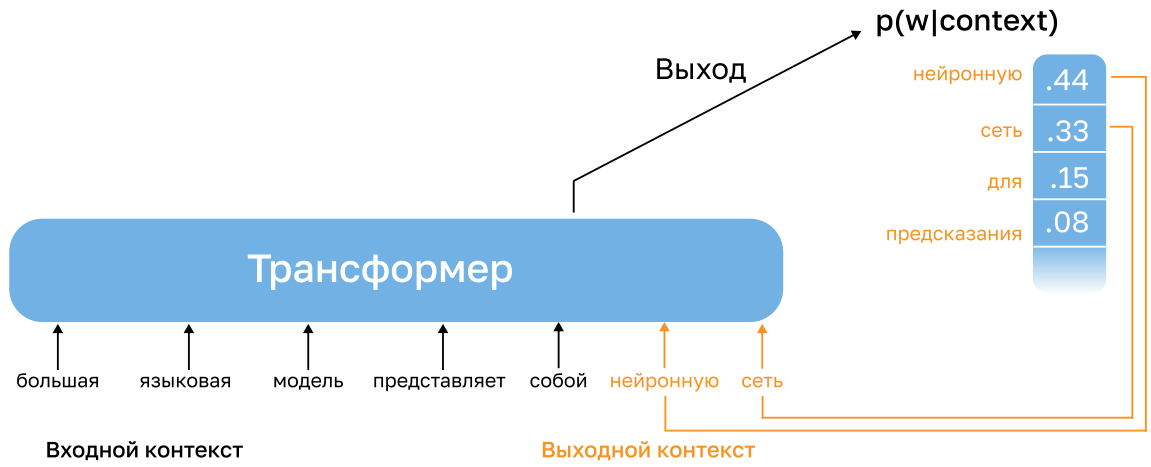
> Синонимы: БЯМ, LLM

Большая языковая модель (Large Language Model — LLM) — это разновидность моделей машинного обучения, которая обучается на огромных массивах текстовых данных, таких как книги, веб-страницы и статьи, и генерирует текст на естественном языке с использованием технологий глубокого обучения. После обучения LLM могут быть адаптированы для решения множества задач, таких как обработка естественного языка и анализ текста, машинный перевод, реферирование статей, написание и отладка кода, генерация естественноязыкового контента.

В основе построения LLM лежит нейронная сеть с архитектурой типа трансформер. Она использует механизм, известный как самовнимание, который позволяет модели учитывать взаимосвязи между словами (токенами) независимо от их положения в тексте. Трансформеры, в отличие от более ранних нейросетевых архитектур (таких как рекуррентные нейронные сети, которые обрабатывают текст последовательно), могут эффективно анализировать сложные шаблоны в больших массивах данных, что позволяет модели лучше учитывать контекст.

Перед обучением LLM текстовые данные (в лингвистике их совокупность обычно называют корпусом текстов) разбиваются на токены, которые затем преобразуются во внутренние векторные представления — **эмбеддинги**. Эмбеддинги позволяют модели учитывать семантические связи между словами и другими языковыми единицами. Благодаря этому LLM при генерации текста может учитывать смысловой контекст и создавать связные, логически последовательные ответы.

Говоря простыми словами, LLM — это вычислительная система, которая предсказывает следующий токен в тексте на основе некоторой совокупности предыдущих слов. Благодаря обучению на огромных объемах данных и большому числу параметров такие модели способны учитывать широкий контекст, генерировать связные тексты и решать разнообразные языковые задачи.



Как видно на рисунке, модель выбирает слово «нейронную», поскольку оно имеет наибольшую вероятность для заданного контекста $p(w \mid context) = 0.44$, подставляет его и затем определяет наиболее вероятное следующее слово для обновленного контекста. Этот процесс повторяется итеративно.

Разработка и применение LLM обычно включают несколько этапов:

- **предобучение (pre-training)** — модель обучается последовательно предсказывать следующие токены на больших массивах текста. Данный этап обучения является наиболее ресурсозатратным;
- **тонкая настройка (fine-tuning)** или дообучение — обучение с учителем на большом корпусе текста, который содержит специально разработанные инструкции языковой модели для понимания языка и ответа на вопросы. На этом этапе LLM обучается решению конкретных задач: отвечать на вопросы, давать резюме, писать код, переводить предложения;
- **обучение в контексте (in-context learning)** — это метод, при котором демонстрация задач интегрируется в подсказку в формате естественного языка. Такой подход позволяет предварительно обученным LLM-моделям решать новые задачи без изменения параметров модели.

Использование LLM связано с рядом проблем, среди которых склонность моделей к галлюцинациям, недостаточная точность и релевантность ответов в отдельных случаях, этические вопросы, низкая объяснимость результатов и высокая стоимость обучения моделей.

Одним из наиболее перспективных направлений применения LLM является их встраивание в процессы управления бизнесом. Наибольшую ценность для бизнеса большие языковые модели приобретают тогда, когда становятся частью аналитических и операционных процессов.

В платформе Loginom для интеграции LLM в конвейеры обработки данных используется библиотека компонентов LLM Kit, которая позволяет применять большие языковые модели непосредственно в аналитических сценариях. Для работы с внешним корпоративным контекстом могут использоваться механизмы RAG и MCP. Подробнее о практическом использовании AI в бизнес-процессах — в разделе «AI в Loginom».