Тема 2

Как работают LLM

Давайте вместе разберёмся, что вообще происходит, когда вы нажимаете «Отправить». Когда вы открываете ChatGPT или Claude.ai, вы пользуетесь *продуктом*, который использует модель — эта страница про саму модель, про «движок» внутри. Без инженерных деталей — ровно столько, чтобы понимать, *почему* ИИ делает то, что делает. Всё остальное в курсе строится на этом.

01 LLM — это не поисковик

Поисковик ищет среди уже существующих документов и возвращает те, что лучше всего подходят под ваш запрос — ничего нового он не создаёт. LLM работает совершенно иначе. Она не выбирает готовый документ, а генерирует новый текст в ответ на запрос. Главное отличие: поисковик возвращает то, что уже существует; LLM создаёт новые предложения.

02 LLM — это «угадыватель» паттернов: она генерирует слово за словом

Вы вводите промт. Модель предсказывает наиболее вероятное следующее слово. Потом следующее. Потом ещё одно. Она генерирует продолжение, которое укладывается в паттерны, которые она выучила. Думайте о ней как об очень способном автодополнении, которое заканчивает не одно слово, а целые абзацы.

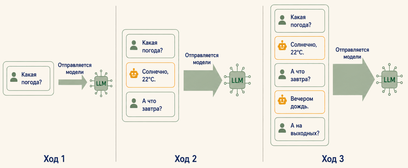
(На самом деле модель генерирует один токен за раз — кусочек текста, обычно это целое слово, иногда его часть. Для простоты будем говорить «слово за словом».)

03 Она училась, читая огромные объёмы текста

Модель обучили на огромном количестве текста: веб-страницах, книгах, статьях, коде и других источниках. Цель обучения — не сохранить копии этих текстов, а выучить закономерности языка: какие слова идут за какими, как строятся предложения, как выглядит перевод, как пишется код. Таких связей — триллионы. Именно поэтому она так хорошо пишет, пересказывает, переводит, кодит и общается с вами — всё это вытекает из одной способности: предсказывать наиболее вероятное следующее слово.

04 У неё нет памяти между сообщениями — она каждый раз перечитывает весь диалог

Когда вы общаетесь с приложением вроде ChatGPT, кажется, что это разговор, правда? Вы спрашиваете — она отвечает, вы пишете в ответ — она помнит. Только модель не помнит. Каждый раз, когда вы нажимаете «Отправить», вся ваша переписка отправляется модели заново — с самого первого сообщения — и она перечитывает всё с нуля. Никакой «памяти о вас» между сообщениями нет. Есть только текст, который каждый раз обрабатывается заново.



05 Она может видеть только ограниченный объём текста за раз

Хотя модель и перечитывает весь ваш диалог каждый раз, у неё есть предел того, сколько она может прочитать за один раз. Этот предел называется *окно контекста*. В длинных чатах ранние сообщения могут выпасть, и модель больше не может их видеть. Когда это происходит, она вдруг как будто забывает, о чём вы говорили, противоречит сказанному ранее или просит объяснить заново.

06 Она может уверенно ошибаться — это следствие принципа работы

Терять контекст — это один способ ввести вас в заблуждение. Вот другой, более коварный: когда модель чего-то не знает — какой-то факт, цитату, сегодняшнюю новость — она может не сказать «я не знаю», а сгенерировать правдоподобный ответ. Выдуманные имена, неверные даты, несуществующие книги или сочинённая статистика могут звучать так же уверенно, как правильный ответ. Это называется *галлюцинация* — и это не баг. Угадывание следующего слова даёт связный текст независимо от того, реальны ли факты под ним.