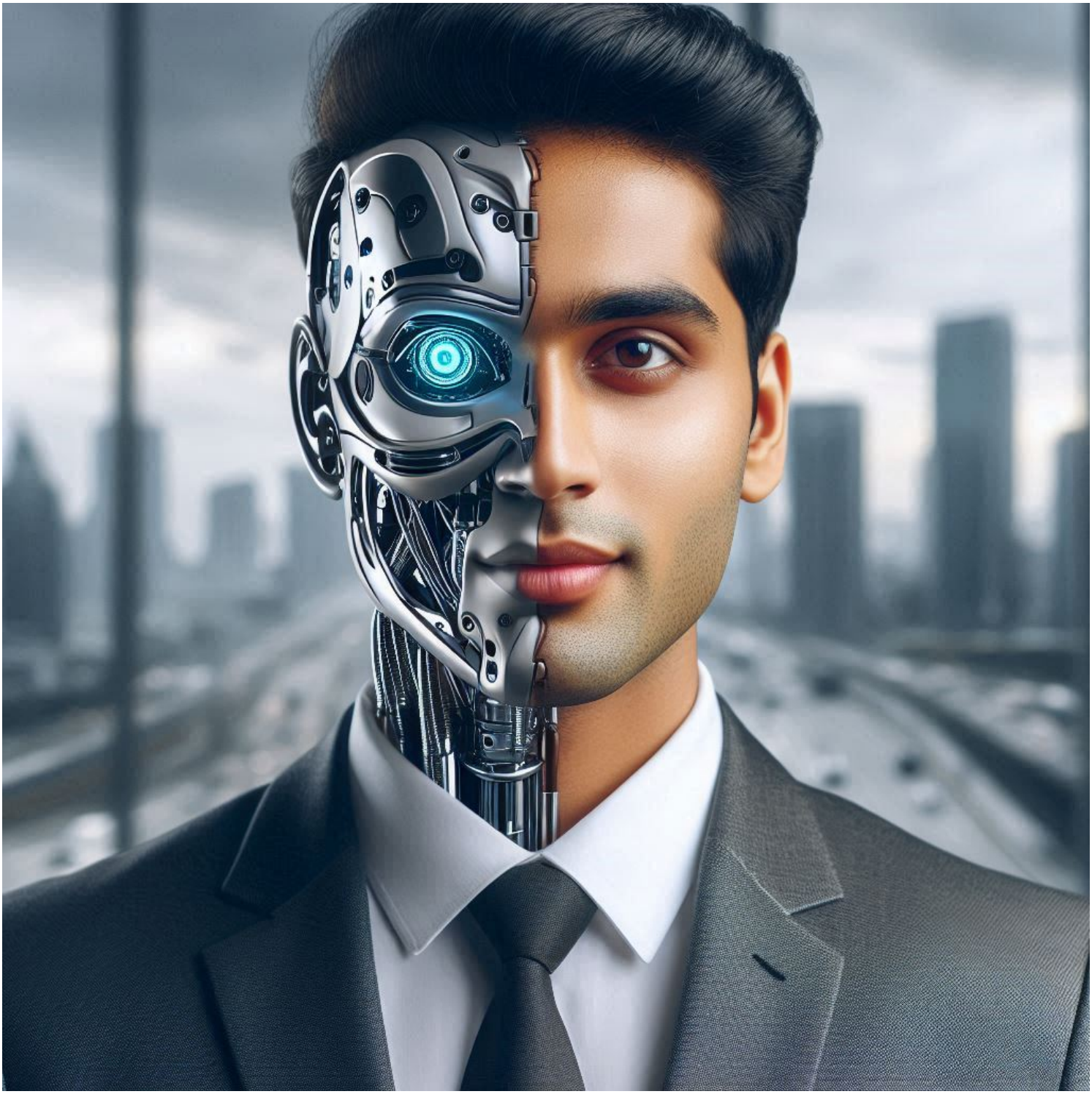
№ 08 (134) 2024 LEGAL TECH

Бионический юрист

10.10.2024 СТАТЬИ LEGAL TECH



Большинство юристов весьма поверхностно понимает возможности искусственного интеллекта (далее — ИИ), особенно его генеративной разновидности. Работать с ИИ без этого знания, конечно, можно, но, если вы не хотите, чтобы вам продали продукт, в котором в GPTs на базе ChatGPT залили Трудовой кодекс и назвали это ИИ-ассистентом по трудовому праву (что сейчас не редкость), необходимо ознакомиться с базовыми принципами его работы, считает Хольгер Цшайге. В этой статье он объясняет, что же действительно может ИИ на

сегодняшний день.

Скорость развития генеративного ИИ

Возможности больших языковых моделей постоянно растут. Это видно по техническим параметрам и результатам выполняемых задач.

Лимит токенов на английском языке *(общий объем обрабатываемой входящей и исходящей информации при запросе к модели)* [1].

Роман «Война и мир» — 680 тыс. слов.
2022 г. — GPT 3.5 4096 токенов — 3000 слов.
2024 г. — Gemini 1.5 Pro 1 млн токенов — 700 000 слов.

Расширение возможностей моделей особенно заметно при сравнении генераторов картинок: на левой — результат первой версии Midjourney, на правой — то, что сейчас умеют генерировать модели (см. фото справа). Видео еще убедительнее [2].

Однако это означает лишь то, что модели становятся все лучше применительно к решению тех задач, которые они технологически способны выполнить. Правда, перечень таких задач становится длиннее. До появления генеративного ИИ существовало довольно жесткое разграничение (то, что легко человеку, сложно ИИ, и наоборот), теперь границы между человеком и алгоритмом частично стираются. И все же период сильного ИИ (artificial general intelligence, AGI [3]) еще не наступил, на пути к его наступлению имеются препятствия.

Факторы, влияющие на скорость развития ИИ

МАТЕМАТИКА Все модели ИИ математические, а прорывы в математике случаются не настолько часто, чтобы за считанные годы многократно ускорить развитие ИИ. С притоком больших денег в эту индустрию стало проще привлекать таланты, но для гениальных

математиков доказательство «Гипотезы Римана» все же на порядок интереснее обучения нейросетей.

КОМПЬЮТЕРНЫЕ МОЩНОСТИ Быстрые GPU (graphics processing unit) нужны не только для компьютерных игр или майнинга криптовалют, но и для обучения моделей ИИ. Возможность параллельного выполнения ими множества вычислительных операций делает их самыми подходящими чипами для компьютерной инфраструктуры ИИ. Нехватка чипов и компьютерных мощностей в целом (так называемый compute) делает обучение больших языковых моделей затратным. OpenAI предлагает услугу обучения кастомных LLM, стоимость которой начинается с $2–3 млн. Говорят, Google потратил на обучение Gemini Ultra $191 млн [4].

ДАННЫЕ По оценкам экспертов, самый острой проблемой для ускоренного развития ИИ на сегодняшний день является доступность обучающих данных. Для обучения больших языковых моделей нужны огромные массивы качественных текстов. Большинство текстов из открытых источников уже использовано для обучения ИИ. Владельцы крупных социальных сетей берут контент пользователей, которые об этом чаще всего даже не догадываются. На очереди публичные форумы и защищенный авторским правом материал — разработчики моделей ведут переговоры с владельцами этого контента. Качество такой текстовой информации может существенно влиять на качество модели.

Как работают большие языковые модели

В многочисленных статьях и вебинарах на тему применения ИИ в работе юристов отмечается их слабое владение понятийным аппаратом из сферы ИИ и отождествление ИИ с нейронными сетями и большими языковыми моделями. А это разные вещи, соотношение которых показано на рисунке.

Нейронные сети часто описывают как компьютерные модели с архитектурой, копирующей архитектуру человеческого мозга: между узлами (нейронами) устанавливаются

определенные связи. Обратим внимание на отличие искусственных нейронных сетей (artificial neural networks, ANN) от естественной сети. Самая большая на сегодняшний день искусственная нейронная сеть состоит из 1,15 млрд нейронов и развертывается на суперкомпьютерах в Sandia National Laboratories [5]. Количество нейронов ANN быстро растет: шесть лет назад самая крупная искусственная нейронная сеть состояла всего из 16 млн нейронов, что сравнимо с размером мозга лягушки, но нынешнее число все равно гораздо меньше человеческого мозга с 86 млрд нейронов. Более того, человеческий мозг управляется не только электричеством, как ANN, но и химическим слоем. Помимо нейромедиаторов (адреналина, дофамина, серотонина и пр.) или опиатов (эндорфинов) собственного производства на его деятельность влияют еще и природные вещества, такие как кофеин или псилоцибин.

Функционал человеческого мозга существенно отличается от функционала искусственных нейронных сетей. У него было два миллиона лет для развития до сегодняшнего состояния. Неслучайно мозг — самая комплексная структура в известной нам Вселенной, благодаря ему мы стали хищником No 1 на Земле. И работает он очень эффективно, в частности по энергопотреблению. Для работы мозгу условно достаточно бутерброда с яичницей на завтрак, тогда когда искусственные нейронные сети потребляют эквивалент энергии атомной станции. Для обучения модели с 100 млрд параметров (стандартная сегодня модель) понадобится примерно 1300 Mwh, а для обслуживания 1 млн пользователей даже маленькой модели (7 млрд параметров) — минимум 55 Mwh [6].

Искусственные нейронные сети являются компьютерными моделями в области ИИ. Большая языковая модель — одна из многих искусственных нейронных сетей. Модели обработки естественного языка (natural language processing, NLP) в ИИ разрабатываются давно, NLP — одно из главных направлений развития ИИ [7]. Среди специалистов большие языковые модели стали популярны в 2017 г. с появлением архитектуры трансформеров [8], которые хорошо подходят для текстов.

Принцип генерации текстов большими языковыми моделями

Рассмотрим самую известную модель трансформеров — генеративный предобученный трансформер (GPT [9]) и покажем, как она работает, вернее, как она генерирует текст (как работают эти модели, никто не знает, даже для своих создателей они являются «черным ящиком», настолько сложна их внутренняя структура). Но никто не знает и то, как работает человеческий мозг.

Поскольку математические модели могут работать только с цифрами, сначала текст надо превратить в цифры [10]. Эта техника называется векторизацией [11]. Первый шаг — превращение текста в так называемые токены, которые отличаются от слов. Для простоты можно использовать токены и слова как синонимы [12].

Простое предложение на английском «I love you!» после токенизации становится таким: «I love you!» — получаются четыре токена (отмечены цветом). Как видите, токены не всегда совпадают со словами, включают пробелы, а знаки препинания превращаются в отдельные токены. Каждому токену присваивается уникальный идентификатор из двух чисел для дальнейшей идентификации, поскольку токен (слово) может иметь несколько векторов в зависимости от контекста и его связи с другими токенами (словами). Например, слово «кот» может иметь отношение к слову «кошка», а также к слову «собака» и даже к слову «сапоги». Все эти отношения выражаются в разных векторах. Представьте себе векторы как обозначение токена (слова) в большом пространстве токенов (слов). Как ширина и долгота на карте обозначают место города, так и векторы обозначают место токена в пространстве токенов. В зависимости от модели это векторное пространство может быть n-мерным. Нам, с нашим представлением трехмерного пространства, практически невозможно представить 12 288-мерное пространство векторов в модели GPT 3.5. Другими словами, каждое слово в этой модели представлено 12 288 векторами.

С превращенными в векторы токенами модель начинает работать. Она анализирует токен

за токеном входящей информации (обычно это называют «промпт»). Главная задача — понять контекст токена и определить вероятность различных последующих токенов, исходя из этого контекста. Токен с наибольшей вероятностью выбирается как последующий после текущего. Человек понимает контекст интуитивно. Например, «Животное переплыло озеро. Оно сильно устало». Понятно, что устало животное, а не озеро. В отличие от человека большая языковая модель не понимает контекст и каждый раз должна пройти многоуровневой процесс. Выглядит это следующим образом.

У архитектуры трансформеров есть так называемые слои [13], через которые проходит последовательный анализ токенов в двухэтапном процессе. На первом этапе (attention step) модель «смотрит» вокруг токена [14], чтобы найти определенный контекст, который нужен ей для определения вероятности разных потенциальных последовательных токенов. На втором этапе (feed forward) вся эта информация перерабатывается, текущий слой определяется с кандидатом на самый вероятный последующий токен, который передается на следующий слой для дальнейшей обработки других возможных контекстов. И так до последнего слоя, в котором модель уже точно определяется с последующим токеном и начинает определять следующий. В итоге данного процесса получается цепочка следующих друг за другом токенов, определенных по наиболее высокой вероятности в зависимости от контекста предыдущих токенов. Эта ваш сгенерированный моделью текст.

Когда модель прекращает генерировать текст

Технологически она должна генерировать бесконечно, поскольку не думает и не понимает смыл сгенерированного текста. Самый простой вариант — модель доходит до собственного лимита токенов. Но чем больше лимиты токенов современных моделей, тем реже возникает такая ситуация. Поэтому разработчики либо задают искусственный лимит объема ответа (не очень точный метод), либо включают в обучающие данные так называемый end-of-sequence token на стадии тонкой настройки модели (более аккуратный метод). Такой токен указывает точку завершения последовательности и помогает модели понять границы между

различными фрагментами текста. Пользователь тоже может ограничить выдачу текста через соответствующий промптинг и изменение параметра «температура». Температура определяет степень вариативности результатов. Чем она выше, тем менее детерминированными являются результаты. Если вы хотите, чтобы модель строго выбрала следующий токен с наибольшей вероятностью (чтобы быть ближе к фактам), то выставляйте низкую температуру модели, если предпочитаете больше креативности в ответе — высокую.

Как обучают языковые модели

Исходной точкой для обучения больших языковых моделей является огромный массив текстовой информации [15]: книги, статьи, вебсайты, программный код и транскрипты аудио- и видеофайлов. Сначала вся эта информация очищается от ошибок. Версия GPT-3 была обучена на текстовой информации объемом 500 млрд слов (для сравнения: десятилетний ребенок сталкивается примерно со 100 млн слов).

Необученную большую языковую модель можно представить как новое музыкальное оборудование, на котором все эквалайзеры на нулевой отметке. Вы слушаете тысячи песен и настраиваете значения эквалайзеров по своему вкусу, пока не добьетесь идеального звука. То же самое происходит с моделями, только в первой итерации настройкой занимаются они сами, математически оценивая отношения слов в текстах и сохраняя эти данные.

В процессе обучения определяются два основных параметра модели: веса (weights) и смещения (biases). Веса — это числовые значения, присваиваемые каждому входному параметру в модели. Вес, связанный с признаком, указывает на его относительную важность для итогового текста модели. Веса используются для линейного объединения входных признаков для получения прогноза модели. Во время обучения модель корректирует веса, чтобы минимизировать разницу между прогнозируемым выходным значением и фактическими целевыми значениями. Смещения — это постоянные значения, добавляемые к взвешенной сумме входных признаков. Они действуют как смещение или перехват в границе принятия решений модели. Чтобы лучше понять, почему определенная модель

генерирует те или иные тексты, важно знать веса и смещения модели. Модели open source публикуют веса и смещения и потому они предпочтительны. Рассмотрим их на примере работы нейронной сети с одним нейроном.

Допустим, вы собираетесь заняться серфингом. Для определения дальнейших действий — идти на пляж или нет — важны три параметра: хорошая погода, температура воды и отсутствие людей на пляже. Сначала определим значения параметров:

погода хорошая, так что x = 1;
вода теплая, так что x = 1;
на пляже много людей, так что x = 0.

Далее каждому параметру придадим вес по шкале от 1 до 5:

погода — 5 (погода важна для серфинга);
вода — 2 (у вас есть костюм, так что температура воды не так уж и важна);
наличие людей на пляже — 3 (несколько человек — не проблема).

Добавим параметр смещения (учитываем предвзятость модели по отношению к определенным результатам) и поставим его значение «-2». Порогом, то есть величиной суммы параметров с весами и смещениями, при которой нейрон «включается», будет 4. Итак, получаем уравнение:

y=5x1+2x1+0x3–2

Результат 5 больше заданного порога 4, так что нейрон включается, значит, мы идем на пляж.

Для большей аккуратности результаты модели после автоматической настройки проходят еще и ручную. В итоге большие языковые модели содержат миллиарды параметров [16]. Чем больше параметров в модели, тем она мощнее.

Как общаться с большими языковыми моделями

Теперь вы знаете, как работают большие языковые модели, и понимаете: чем больше и подробнее входящая информация (промпт), тем выше вероятность качественного результата. Вокруг правильного промптинга образовалась новая профессия — prompt engineers, это деятельность профессионалов, умеющих грамотно озадачивать модели для получения максимально четкого и полезного результата. В настоящее время спрос на этих специалистов высок и зарплата у них заоблачная (за навык грамотно формулировать запрос алгоритму). Однако уже в скором будущем ситуация изменится: модели будут брать на себя интерпретацию запросов пользователей и превращать их в соответствующий промпт. Большие языковые модели пойдут по пути Google и Яндекса и будут использовать ИИ для оптимизации запросов.

Стратегии составления промптов

INTENT + CONTEXT + INSTRUCTION

Четко сформулируйте желаемый результат, предоставьте соответствующую справочную информацию и укажите, какие действия необходимо совершить.

Пример: Проанализируйте положения договора, касающиеся прав интеллектуальной собственности в контексте спора о лицензировании программного обеспечения. Составьте возможный встречный иск, основанный на нарушении авторских прав.

ROLE PLAYING

Назначьте LLM определенную роль.

Пример: Вы — опытный судебный юрист. Разработайте стратегию перекрестного допроса для ключевого свидетеля в деле о врачебной халатности.

CONDITIONING

Установите параметры или ограничения для ответа LLM.

Пример: Составьте проект мирового соглашения, предполагая, что истец требует возмещения ущерба в размере 1 млн рублей, но готов согласиться на 500 тыс. рублей.

CHAIN-OF-THOUGHT PROMPTING

Позвольте LLM разбивать сложные проблемы на более мелкие этапы.

Пример: Проанализируйте элементы иска о нарушении контракта. Определите, подтверждают ли факты дела каждый элемент.

SYSTEM PROMPTING

Предоставьте инструкции или рекомендации, которые повлияют на общее поведение LLM.

Пример: Сосредоточьтесь на предоставлении кратких и действенных юридических консультаций.

Лайфхаки промптинга для улучшения результатов

1 Поставьте инструкцию в начало промпта и отделите от контекста символами ### или """.

Пример: *Обобщите текст в виде списка самых важных терминов.*

Текст: ###
{Вставьте текст здесь.}
###

2 Будьте конкретны, описательны и максимально подробны в отношении желаемого контекста, результата, длины, формата, стиля и пр.

Пример: *Напишите вдохновляющее стихотворение о сложности соблюдения правил GDPR в стиле Шекспира.*

3 Сформулируйте желаемый формат вывода с помощью примеров.

Пример: *Извлеките важные сущности, упомянутые в тексте: сначала — все названия компаний, затем — все имена людей, далее — темы, которые соответствуют содержанию, и, наконец, — общие темы.*

> Желаемый формат:
> Названия компаний: <comma_separated_list_of_ company_names>
> Имена людей: -||-
> Конкретные темы: -||-
> Общие темы: -||-
> Текст: {text}

В промптинге есть много более сложных подходов для улучшения результатов. Стратегии и подходы в нем часто объединяют во фреймворки. Для ясности концепции приведем один фреймворк:

R-T-F: Role, Task, Format (Роль, Задача, Формат). Действуй как [РОЛЬ].
Создай [ЗАДАЧА].
Покажи, как [ФОРМАТ].

Пример: *Действуй как [младший юрист]. Создай [обобщение] приложенного судебного решения. Покажи в виде [списка].*

А что с «галлюцинациями»

Возможность больших языковых моделей генерировать бессмысленную информацию часто называют главной причиной их неготовности к использованию в профессиональной среде. Слишком велик риск ошибок. При этом к данному явлению относятся, как к ошибке моделей, но это не ошибка, а их свойство. Главная задача моделей — генерировать текст на основе большого массива текстовой информации, на котором их обучали. При этом модели не озадачиваются вопросом по поводу того, имеет ли смысл сгенерированное ими. У них нет возможности проверить это. Другими словами, большие языковые модели не думают, как это делает человек.

Относитесь к генеративному ИИ, как к юристу, которого учили никогда, ни при каких обстоятельствах не говорить: «Я не знаю». Надо дать любой ответ. Есть разные способы борьбы с этой проблемой, например повысить порог вероятности следующего токена. Условно модель определила вероятность пяти потенциальных кандидатов на следующий токен из контекста анализа предыдущих. Но самая большая вероятность из пяти составляет 63%, а в качестве порогового значения установлено 75%. В таком случае модель не продолжает генерировать текст, а отвечает: «Я не могу дать ответ».

В настоящее время самым популярным методом ограничения «галлюцинаций» является retrieval-augmented generation (RAG). Результат модели сверяется с базой проверенной информации, например со справочной правовой системой (СПС). Если модель выдала судебное решение, которого в СПС нет, значит, она его придумала [17]. Вендоры продуктов

на базе генеративного ИИ заявляют, что путем применения RAG и обучения на проверенной информации им удалось свести «галлюцинации» своих моделей к нулю, но на практике определенный процент таковых все равно остается.

Это не делает модели совершенно бесполезными. В задачах, где аккуратность информации критична, надо ввести дополнительный шаг проверки. И не стоит забывать, что живые юристы тоже ошибаются (по разным исследованиям, даже чаще, чем алгоритмы). К сожалению, человек требует от алгоритма стопроцентной аккуратности в работе, в то время как к себе гораздо менее требователен.

Какие еще пробелы наблюдаются в генеративном ИИ? Перечислим основные.

ДЛИНА КОНТЕКСТА Не все модели обладают необходимым лимитом токенов для того, чтобы одновременно перерабатывать большой объем информации. Например, если вы загружаете в модель большой документ и просите обобщить его, то она вполне может «забыть» начало при анализе последних страниц.

СТОИМОСТЬ Операционные расходы на создание и поддержку больших языковых моделей еще достаточно высоки для того, чтобы оправдать большое количество запросов. Например, если вы загружаете в модель 200-страничный договор и направляете 150 вопросов, то такой анализ договора может обойтись дороже работы

ИЗМЕНЧИВОСТЬ ПРОМПТОВ Большие языковые модели чрезвычайно буквальны и чувствительны к промптам. Иногда требуется большое количество итераций промпта для получения адекватного результата. Также модели не обладают логикой и не могут сделать выводы из информации, которой их обучали. Например, если спросить ChatGPT, кто мама Тома Круза, модель правильно ответит, что это Мэри Ли Пфайффер, но, если спросить, кто сын Мэри Ли Пфайффер, она не найдет правильного ответа.

ПЕРЕКРЕСТНЫЕ ССЫЛКИ И ОПРЕДЕЛЕНИЯ У моделей есть проблемы с перекрестными

ссылками на ранее обработанный текст, поскольку повторно они его не «читают».

СИНОНИМЫ Модели испытывают сложности со словами, которые в общем языке не являются синонимами, зато являются таковыми в юридическом языке. Это связано с тем, что общие модели были обучены не на юридических текстах.

ЮРИДИЧЕСКИЙ ЖАРГОН У больших языковых моделей есть трудности с юридическим языком.

НОРМАЛИЗАЦИЯ [18] И ОБЪЯСНИМОСТЬ Проблемой является обеспечение последовательности и объяснимости ответов (последнее — особенно актуально для закрытых моделей) [19].

Применение больших языковых моделей на практике

Итак, мы дошли до главного вопроса: как применять большие языковые модели на практике? Подавляющая часть работы юристов связана с текстами, поэтому модели, которые их генерируют, должны быть идеально адаптированы к работе данных специалистов. Глядя на кейсы применения, ИИ можно разделить на детерминистический и вероятностный (так называемый генеративный), а работу юристов — на юридическую и административную. Начнем с юридической работы, поскольку здесь можно применять и детерминистический, и вероятностный ИИ.

Детерминистический ИИ

Экспертные системы. Экспертные системы (ЭС) — это первые варианты детерминистического ИИ (ИИ с заданным объемом информации), которые не могут генерировать новую информацию, а способны только воспроизводить информацию в модели. ЭС успешно применяются юристами уже 40 лет.

Извлечение данных и категоризация. Уже более 15 лет ИИ используется для извлечения данных и категоризации, например договоров. Если вам нужно проанализировать большой объем информации и извлечь ключевые данные (в рамках eDiscovery или Due Diligence), ИИ существенно сократит срок и стоимость работы.

Предиктивная аналитика. ИИ и до появления генеративного варианта хорошо подошел к выявлению характеристик (pattern recognition) в большом объеме данных, что позволило делать определенные прогнозы. Разные решения использовали это свойство ИИ для предиктивного анализа судебной практики.

Вероятностный (генеративный) ИИ

Поиск, анализ и обобщение юридической информации. Генеративный ИИ отлично справится с детализированным поиском юридической информации: не только более точно подберет законы, судебную практику, статьи и другую текстовую информацию, но и сможет в определенной степени проанализировать и обобщить ее.

eDiscovery. Так же, как и с поиском общей юридической информации, генеративный ИИ лучше справляется с выявлением конкретной информации из большого массива данных, документов и пр.

Предиктивная аналитика. Генеративный ИИ лучше детерминистического выявляет характеристики в большом массиве данных и на основе этого составляет прогноз.

Создание документов. Вероятностные модели могут генерировать новую текстовую информацию, их хорошо использовать для создания юридических документов (договоров, исков, меморандумов и пр.). Юристам, конечно, следует проверять такие документы, но их автоматическая генерация значительно экономит время на создание таковых.

Анализ договоров. Генеративные модели могут анализировать договоры на предмет отклонения от принятых условий, ошибок и пр. В этом они превосходят детерминистические

модели.

Управление знаниями. Обычно знания хранятся в текстовой форме, поэтому генеративный ИИ хорошо подходит для их систематизации и поиска в системе управления таковыми.

Due diligence. Генеративный ИИ может существенно сократить время на анализ информации и выявление рисков в рамках крупных трансакций.

Мониторинг комплаенса. Усложнение регуляторной среды — одним из главных вызовов для юристов. Вручную контролировать соответствие деятельности компании нормам и правилам практически невозможно. Генеративные модели могут брать на себя автоматизированный комплаенс-контроль.

B2C сервисы / ИИ-агенты. Большая доля юридических запросов населения и малого бизнеса является типовой и может быть автоматизирована. На основе генеративных моделей можно создать платформы и ИИ-агентов, которые превратят услуги в продукты и существенно снизят стоимость юридических услуг, упростив доступ к ним.

На сегодняшнем этапе развития большие языковые модели больше подходят для операционной, а не юридической работы, потому что риск ошибиться при выполнении операционной работы существенно ниже и в целом может быть проигнорирован. Уже сегодня модели могут управлять электронной почтой, заниматься онбордингом новых клиентов, автоматически учитывать потраченное на проекты время и выставлять счета, отвечать на вопросы сотрудников и клиентов, придумывать и реализовать стратегии развития бизнеса. Это лишь небольшой список задач, для выполнения которых пригодятся большие языковые модели. Еще несколько десятков кейсов применения могут найти сами модели.

Российские LegalTech решения, использующие ИИ

- платформы для экспертных систем — botman.one;
- анализ контрактов — Embedika Contract, ABBYY Compreno (теперь InfoExtractor SDK), Noroots;
- предиктивная аналитика — «Сутяжник», Casebook;
- управление интеллектуальной собственностью — PatentCore;
- ИИ-агенты — Doczilla, Pravo(tech);
- сервисы B2C / B2B — «Правовед», «Европейская юридическая служба».

Будущее юристов

Потенциал ИИ порождает у юристов обеспокоенность: не вытеснит ли он их с рынка юридических услуг? Понять юристов можно, но дискуссия на эту тему контрпродуктивна. При всем огромном потенциале ИИ не стоит забывать, что это инструмент. Безусловно, данный инструмент возьмет на себя часть той работы, которую сегодня делают юристы, и их работа будет меняться так же, как и операционная модель, а в итоге и бизнес-модель рынка юридических услуг.

40 лет назад многие прогнозировали смерть профессии бухгалтера в связи с появлением Lotus 1–2–3 (а позже — Excel). И действительно, простых счетоводов стало меньше, зато существенно выросло число финансовых аналитиков, аудиторов и прочих специалистов, работающих с финансовыми данными. Аналогичное развитие можно прогнозировать и для юридического рынка труда. Действует парадокс Джевонса: технологический прогресс, повышая эффективность использования какого-либо ресурса, увеличивает, а не уменьшает объем его потребления. В данном случае ресурс — работа юристов. Другими словами, работы для юристов будет не меньше, а даже больше, гораздо больше.

Прообразом юриста будущего скорее всего является бионический юрист — симбиоз живого юриста и технологий. Это новая единица, части которой не могут эффективно действовать отдельно друг от друга. Илон Маск в разговоре с Лексом Фридманом обратил внимание на такой аспект, как скорость коммуникации. У человека она составляет примерно 1 bps [20], у ИИ — свыше 1 Gbps. Представьте себе, как быстро при такой скорости два ИИ смогут согласовывать между собой условия договора. При этом условия по-прежнему будут задавать люди, ИИ просто избавит их от переписки и телефонных разговоров.

---

[1] Количество токенов за слово варьируется в зависимости от языка.

[2] https://youtu.be/QRuFtMNCta8?si=4F6H-fO9hKCTd_x9

[3]

То, что вы увидели в «Матрице» и «Терминаторе».

[4] 2024 AI Index Report. — https://aiindex.stanford.edu/report/

[5] Крупный НИИ американского ВПК.

[6] https://adasci.org/how-much-energy-do-llms-consume-unveiling-the-power-behind-ai/

[7] Настоящий фурор в свое время вызвала модель ELIZA 1967 г. А знаменитый тест Тьюринга, предложенный в 1940 г., оценивает способность алгоритма анализировать человеческий язык.

[8] Трансформеры являются разработкой Google.

[9] Для большей ясности: OpenAI просто выбрал для своей LLM название данного рода моделей.

[10] Все алгоритмы работают с цифрами, на базовом уровне — с «0» и «1».

[11] Под названием «Word2vec» была разработана в Google в 2013 г.

[12] Количество токенов за одинаковые слова на разных языках сильно варьируется. Так, на текст на венгерском уходит в 16 раз больше токенов, чем на тот же текст на английском. Это важно знать, чтобы учитывать лимит токенов моделей при работе с иностранными текстами.

[13] У GPT-3.5 96 слоев, у GPT-4-120. Чем больше количество слоев, тем точнее результаты.

[14] На все предыдущие обработанные токены.

[15] GPT был обучен на датасете Common Crawl (https://commoncrawl.org/).

[16] GPT-4 содержит 1,8 трлн параметров.

[17] Разумеется, если в СПС содержатся все судебные решения.

[18] Нормализация — это обеспечение последовательности ответов. В зависимости от постановки вопроса модель может выдавать совершенно разные ответы на один и тот же вопрос.

[19] Все модели, по сути, — «черные ящики», поэтому проблема объяснимости актуальна для них для всех, даже для открытых.

[20] Бит за секунду.

Хольгер Цшайге

Генеральный директор "Инфотропик Медиа", Член правления ELTA, Член Advisory Board Global Legal Tech Hub

СТАТЬИ LEGAL TECH