RU EN ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ ISSN 2079-3316
научная статья ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ

УДК 004.89:343.15
10.25209/2079-3316-2026-17-1-21-56

Анализ судебных решений с помощью больших языковых моделей

Юрий Петрович **Сердюк**[1], Наталья Александровна **Власова**[2], Седа Рубеновна **Момот**[3], Елена Анатольевна **Сулейманова**[4]
[1–4] Институт программных систем им. А. К. Айламазяна РАН, Веськово, Россия

Аннотация. В статье рассматривается использование больших языковых моделей последнего поколения (таких как ChatGPT, Grok, DeepSeek, GigaChat, YandexGPT) для анализа судебных решений. В анализе использовались дела гражданской, административной и уголовной категорий.

Составлен датасет текстов судебных решений, взятых из базы судебных и нормативных актов РФ, с официального портала судов общей юрисдикции г. Москвы, с сайта Российского агентства правовой и судебной информации.

Предложено и реализовано несколько видов тестов больших моделей, сформулированы принципы выбора эталонных ответов, а также тексты запросов (промпты). Проверялись возможности больших моделей в прогнозировании апелляционных решений, квалификации преступных деяний, оценке решений нескольких видов инстанций по одному делу. Также проверялась способность моделей к вынесению собственных последовательных решений.

Результаты тестирования показали, что точность прогнозирования судебных решений на реальных случаях с помощью больших языковых моделей лишь в отдельных случаях превышает 50%. Приведен обзор статей по использованию ИИ в юридической практике.

Ключевые слова и фразы: большие языковые модели, БЯМ, судебные решения, датасет, промпт, ИИ в юриспруденции, large language models, LLM, LegalAI

Для цитирования: Сердюк Ю. П., Власова Н. А., Момот С. Р., Сулейманова Е. А. *Анализ судебных решений с помощью больших языковых моделей* // Программные системы: теория и приложения. 2026. **Т. 17**. № 1(70). С. 21–56. https://psta.psiras.ru/read/psta2026_1_21-56.pdf

Введение

С начала 2025 г. начался новый этап в развитии и применении систем искусственного интеллекта (ИИ) – так называемый генеративный (generative) ИИ был расширен возможностями рассуждений, что существенно повысило способности ИИ в решении почти всех интеллектуальных задач: от математических до задач постановки диагнозов в медицине, а также планирования и управления в различных областях экономики. Новые возможности систем ИИ в области анализа и рассуждений позволили расширить сферу применения последних, в том числе в юридической практике (в так называемой области LegalAI). Так, например, если на декабрь 2023 г., согласно опросу, выполненному экспертами журнала LegalInsight [1], только 14,5% юристов в России активно использовали инструменты генеративного ИИ и еще 33,9% планировали их использовать в будущем, то на август 2025 г. таких было уже 28% и 43% соответственно [2]. Таким образом, в экспертном сообществе всё более утверждается мнение, что ИИ в ближайшее время полностью трансформирует сектор юридических услуг [3, 4].

В мире рынок LegalAI в 2024 г. превысил $2 млрд с прогнозом роста до $3,64 млрд к 2030 году при среднегодовом темпе роста около 10,4%. Около 60% инвестиций в LegalAI приходится на долю США, Великобритании и Китая. Согласно исследованию компании WoltersCluwer [5], 74% юридических фирм в Европе и Северной Америке уже внедрили решения на базе ИИ.

Согласно обзору, составленному компанией Gartner [6], ИИ в юридической практике применяется для решения следующих основных задач:

(1) составление и анализ договоров,
(2) поиск и исследования по делам судебной практики,
(3) резюмирование и структурирование документов,
(4) извлечение данных (параметров контрактов, условий, сумм и дат) из документов,
(5) анализ правовых и налоговых рисков в документах и контрактах,
(6) автоматизация документооборота: генерация типовых справок, налоговых деклараций и т.п.

Задачи анализа, исследования и прогнозирования судебных решений требуют от систем ИИ в первую очередь знаний в тех или иных областях права. Соответственно, к данному моменту имеется достаточно много работ, посвящённых созданию датасетов юридической направленности и их использованию для тестирования больших языковых моделей

(БЯМ). В частности, созданы датасеты для оценки способностей БЯМ к рассуждениям в области налогового права РФ [7], общей части уголовного права РФ [8], а также датасеты экзаменационных вопросов по различным видам права [9]. Кроме того, имеются датасеты с описанием судебных дел, предназначенные для тестирования моделей на предмет применения ими различных видов доказательств и их использования в заключительном решении (приговоре) [10].

Область прогнозирования судебных решений LJP (legal judgment prediction) с помощью ИИ является в настоящий момент активной областью изучения и экспериментов. В последнее время появилось большое количество статей, представляющих разные подходы к решению этой задачи, в частности описывающих соответствующие датасеты из судебных решений, принятых в судах различных стран [11–16].

Та часть работ, которые посвящены тестированию существующих моделей на задаче LJP, нередко имеют узкие рамки и часто обладают следующими недостатками:

(1) тестирование БЯМ на специально подобранных, упрощенных или искусственных случаях (например, когда исходные документы, которые анализировались моделями, специально модифицировались с целью выяснения отдельных аспектов судебного дела [15]),
(2) упрощенные требования к результатам прогноза решения (часто задача сводится к простой классификации: «виновен/не виновен», или «подтвердить приговор/отменить приговор/частичное подтверждение или отмена» [12]),
(3) отсутствие обоснования эталонности ответов, с которыми сравниваются ответы моделей,
(4) отсутствие в большинстве работ целостного, комплексного подхода к тестированию БЯМ, при котором проверяется сразу несколько важных свойств и возможностей моделей,
(5) использование для тестирования нейронных сетей предыдущих поколений, не обладающих механизмами рассуждений,
(6) тестирование, в основном, только какой-то одной конкретной модели, что не позволяет выявить лучшие модели для решения определённых юридических задач.

Наконец, в опубликованных исследованиях практически отсутствуют работы по прогнозированию решений российских судов с использованием рассуждающих БЯМ последнего поколения. На данный момент имеются лишь статья из блога «Судебная практика» с сайта zakon.ru [17] и статья с сайта law.ru [18]. В первой из них приведены результаты оценки дела Ларисы Долиной по продаже квартиры, которые были получены

системами ChatGPT-5.1, Claude Opus 4.1, Gemini 2.5 Pro и YandexGPT 5.1 Pro. Во второй статье описан успешный опыт применения системы Gemini для составления апелляционной жалобы, позволившей отменить решение суда первой инстанции.

Отличительными особенностями настоящей работы являются следующие:

(*1*) тестирование больших языковых моделей проводилось на реальных делах трех категорий – гражданских, административных, уголовных; судебные решения (первой инстанции и апелляционные) брались из открытых источников судебной и правовой информации (см. раздел 1);

(*2*) в качестве ответов от БЯМ требовались как краткие ответы (например, «*есть ли основания для пересмотра дела – да/нет*»), так и их объяснения с указанием конкретных статей и пунктов того или иного кодекса, а также некоторые другие сведения;

(*3*) эталонность ответов, с которыми сравнивались ответы моделей, обеспечивалась специальным подбором судебных решений (см. более подробно об этом в разделах 1 и 2);

(*4*) комплексный подход к тестированию БЯМ состоял в проведении нескольких разноплановых тестов:

- (*а*) прогнозирование решений апелляционного суда (раздел 2.1);
- (*б*) проверка моделей на последовательность вынесения ими решений (раздел 2.2);
- (*в*) проверка моделей на совокупности решений нескольких инстанций по одному и тому же делу (раздел 2.3);
- (*г*) квалификация преступных деяний в соответствии с конкретными статьями, их частями и пунктами (уголовного кодекса РФ) (раздел 2.4);

(*5*) тестирование проводилось на больших языковых моделях последнего поколения: *GPT-5*[1], *DeepSeek V.3.2-Exp*[2], *Grok-4*[3], *GigaChat 2.0*[4] и *YandexGPT 5.1 Pro*[5] в режиме рассуждений;

(*6*) все модели проверялись на четырех тестах (см. п. (*4*) выше) и ранжировались по результатам.

Стоит отметить, что при тестировании моделей запросы (промпты) были двух типов:

(*а*) полный текст запроса задавался во входном окне модели,

(*б*) к запросу в окне добавлялись присоединённые файлы.

[1] https://chatgpt.com/
[2] https://chat.deepseek.com/
[3] https://grok.com/chat
[4] https://giga.chat/
[5] https://alice.yandex.ru/

Однако RAG (Retrieval Augmented Generation) – подход, который сочетает использование БЯМ совместно с некоторой внешней базой данных (юридических документов), в данной работе не применялся. Тем не менее, в разделе 3 даётся обзор некоторых работ на эту тему [19].

Также в данном исследовании использовались только простые одношаговые промпты. Оценка многошаговых диалогов с большими моделями затруднена из-за отсутствия на настоящий момент общепринятых критериев правильности ответов моделей в таких диалогах [20].

Статья построена следующим образом. В разделе 1 описан датасет судебных и апелляционных решений, на которых проводилось тестирование моделей. Раздел 2 посвящён отдельным тестам (подразделы 2.1–2.4). В разделе 3 приводится обзор последних работ (в основном 2025 года) по прогнозированию судебных решений с помощью больших моделей. Одновременно даётся сравнение этих работ с исследованием, описанным в настоящей статье. В заключительном разделе 4 дано краткое изложение полученных результатов вместе с выводами и перечислением направлений дальнейшей работы.

Материалы данного исследования: промпты, протоколы с ответами моделей, таблицы результатов и др. *находятся в свободном доступе*[6].

1. Датасет судебных и апелляционных решений

Для проведения тестирования больших моделей на реальных делах из судебной практики был составлен датасет судебных и апелляционных решений, которые были взяты из следующих открытых источников:

(1) *интернет-ресурс «Судебные и нормативные акты РФ»*[7],
(2) *официальный портал общей юрисдикции города Москвы*[8],
(3) *сайт Российского агентства правовой и судебной информации (РАПСИ)*[9],
(4) *сайт Верховного суда РФ*[10].

В состава датасета вошли дела судов разных инстанций, в частности Верховного суда РФ. В общей сложности датасет состоит из более чем 50 судебных дел – как отдельных приговоров судов первой инстанции, так и пар решений судов первой и апелляционной инстанций, а также изложений решений судов нескольких инстанций по одному делу, рассматривавшемуся, в конечном итоге, Верховным судом РФ.

[6] https://cloud.mail.ru/public/a5t4/d83UvenhJ
[7] https://sudact.ru/
[8] https://mos-gorsud.ru/
[9] https://rPSInews.ru/
[10] http://vsrf.ru/

Датасет состоит из трёх частей:

(*1*) судебные дела для прогнозирования решений апелляционного суда и проверки моделей на последовательность в вынесении решений (тесты 1 и 2): 10 административных дел, 13 гражданских и 9 уголовных, всего 32 дела;

(*2*) изложение решений судов нескольких инстанций по одному делу (тест 3): 10 дел различных категорий;

(*3*) судебные дела для квалификации преступных деяний и предсказания статей обвинения (тест 4): 10 уголовных дел.

Все судебные дела брались из открытых источников, поэтому они в основном, но в разной степени являются деперсонифицированными. Например, в судебных делах уголовной категории обычно пропущены фамилии фигурантов, некоторые адреса, но имена и фамилии подсудимых могут быть сохранены. Необходимо заметить, что в некоторых случаях деперсонификация искажает смысл предложения. Кроме того, обезличивание не везде было выполнено последовательно. Например, одно и то же лицо в тексте решения суда в одном месте называется «ФИО1», а в другом по имени и отчеству. Следовательно, при тестировании модели не пользовались всем объёмом информации, который необходим для корректной интерпретации анализируемых дел, что могло отразиться на правильности их ответов.

2. Анализ судебных решений

В данном разделе описаны четыре теста, которые были выполнены для систем GPT-5, DeepSeek, Grok, GigaChat и YandexGPT:

Тест 1 – прогнозирование апелляционных решений,

Тест 2 – проверка на последовательность принимаемых моделями решений (с использованием результатов Теста 1),

Тест 3 – прогнозирование решения Верховного суда РФ по решениям нескольких видов инстанций по одному делу,

Тест 4 – квалификация преступных деяний уголовной категории с прогнозированием статей обвинения.

2.1. Прогнозирование апелляционных решений

Целью первого эксперимента было установить, в какой степени современные большие языковые модели способны прогнозировать решение апелляционной инстанции, опираясь на решение суда первой инстанции. Для верификации ответов моделей использовались реальные апелляционные определения. Соответственно, для этого теста подбирались пары

вида «решение суда первой инстанции + решение суда апелляционной инстанции». Поскольку решения суда апелляционной инстанции выступали в качестве «эталонной» экспертной оценки, они должны быть законными и обоснованными, насколько это вообще возможно для данной предметной области. Для этого при подборе судебных решений мы опирались на следующие критерии:

Критерий 1. С самого начала были исключены случаи, в которых решения апелляционного суда опираются на данные, не отражённые в решении суда первой инстанции (например, стороны достигли соглашения после вынесения решения суда первой инстанции или одна из сторон предоставила документы, которыми не располагал суд первой инстанции). Возможные недостатки решения суда первой инстанции должны были определяться на основании самого текста. По этой причине отбирались только те случаи, в которых апелляционный суд рассматривал возможные неправильное определение или недоказанность обстоятельств, имеющих значение для дела, или несоответствие выводов суда обстоятельствам дела (что соответствует ч. 2 ст. 310 КАС, ч. 1 ст. 330 ГПК, п. 1 ст. 389.15 УПК).

Критерий 2. Отбирались лишь такие случаи, в которых решение апелляционного суда либо не оспаривалось в вышестоящей инстанции, либо оспаривалось, но было оставлено в силе вышестоящей инстанцией.

Для краткости в результатах теста ответы, совпавшие с решением апелляционной инстанции, отмечались как «правильные», а несовпавшие — как «неправильные».

Таким образом было подобрано 32 пары судебных решений в трёх категориях: административные, гражданские и уголовные. В каждой категории есть как оставленные в силе, так и отменённые или изменённые апелляционной инстанцией (таблица 1).

Таблица 1. Распределение дел по категориям в тесте

Категория дела	Решением апелляционной инстанции		Всего
	оставленных в силе	отменённых или изменённых	
Административные	4	6	10
Гражданские	9	4	13
Уголовные	2	7	9
Итого	**15**	**17**	**32**

Суть теста 1 заключалась в том, что по полному решению суда первой инстанции модель должна определить, имеются ли основания для пересмотра дела апелляционной инстанцией, а также какое апелляционное определение должно быть вынесено в соответствии со статьёй о полномочиях апелляционной инстанции. Соответственно, структура промпта для

выполнения этого теста была следующей (полный текст данного промпта приведён в приложении):

(*1*) полный текст статьи процессуального кодекса, содержащей основания отмены или изменения судебного решения в апелляционном порядке (соответственно категории дела это или ст. 310 КАС, или ст. 330 ГПК, или ст. 389.15 УПК);

(*2*) полный текст статьи процессуального кодекса, посвящённой полномочиям суда апелляционной инстанции (соответственно категории дела это или ст. 309 КАС, или ст. 328 ГПК, или ст. 389.20.1 УПК);

(*3*) непосредственно вопросы, на которые должна была ответить модель.

Этими вопросами были следующие.

Вопрос 1. *Есть ли основания для пересмотра дела согласно статье* [номер статьи, содержащей основания отмены или изменения судебного решения, для данной категории дел]*? (отвечай «да» или «нет»). Если «Да», то приведи 1–2 самых главных из оснований.*

Вопрос 2. *Каким должно быть решение суда апелляционной инстанции в соответствии со статьёй* [номер статьи, содержащей полномочия суда апелляционной инстанции, для данной категории дел]*? Приведи номер пункта.*

Вопрос 3. *Если возможны разные варианты по статье* [номер статьи, содержащей полномочия суда апелляционной инстанции, для данной категории дел]*, то приведи вероятность в процентах для каждого варианта.*

В итоговые результаты из ответа модели отбирались:

(*а*) ответ «да» или «нет» на первый вопрос;

(*б*) номер пункта статьи, в соответствии с которым должно быть вынесено решение апелляционной инстанции (такие как «оставить решение суда первой инстанции в силе», «отменить решение суда первой инстанции и назначить новое решение» и т.п.);

(*в*) вероятность в процентах для пункта статьи, совпадающего с определением апелляционной инстанции.

Пункт (*в*) играл вспомогательную роль для более объективной оценки ответов моделей. Например, модели GigaChat и YandexGPT, как правило, давали вероятность 90%–100% для выбранного ответа, тогда как система Grok, кроме двух случаев, давала более умеренную оценку в 60–70% для выбранного варианта.

Необходимо специально отметить следующие важные особенности ответов моделей на вопросы теста.

(*1*) Некоторые модели в ответе на вопрос 1 «*Есть ли основания для пересмотра дела?*» выбирают из двух вариантов ответа («да» или «нет») один и тот же вариант почти для всех дел. Например, система GigaChat для гражданских и административных дел в 18 из 23 случаев ответила, что «оснований для пересмотра нет». Так как этот ответ в отдельных случаях совпадал с принятым за эталонное апелляционное определение, то он засчитывался как «правильный», хотя из-за однообразности ответов модели это могло быть просто случайным попаданием.

Оценить это можно, только анализируя обоснования, которые приводят в этом случае модели, отвечая на вопрос 2 данного теста, требующий предсказать номер пункта статьи, в соответствии с которым должно быть вынесено решение апелляционной инстанции.

(*2*) Даже если ответ модели на вопрос 2 (пункт статьи) совпадал с (эталонным) ответом апелляционного суда, обоснования для этого ответа нередко отличались от тех, которые были приведены в апелляционном определении. В таблице 2 в качестве иллюстрации такой ситуации, приведены фрагменты ответов моделей.

Таблица 2. Фрагменты ответов моделей

	Фрагмент ответа модели Фрагмент решения суда апелляционной инстанции
(1)	YandexGPT: «Нарушение правил о языке судебного производства (п. 3 ч. 4 ст. 330 ГПК РФ). В решении суда упоминается использование иностранных названий компаний («Nano IT» SIA, Layer6 Networks), что может указывать на нарушение требований к языку судопроизводства . . . » « . . . установлен факт нарушения исключительных прав истца на сообщения в эфир телепередач телеканала . . . »
(2)	DeepSeek: « . . . неправильное истолкование Налогового кодекса РФ судом первой инстанции». «Указанный досудебный порядок административным истцом не соблюдён».

В случае с фрагментом (1) очевидно, что предполагаемое основание для отмены решения суда выбрано по поверхностному признаку – использование латиницы, в результате чего сделан вывод о возможном нарушении правил о языке судебного производства. В апелляционном определении по этому делу учтена исключительно суть дела о нарушении интеллектуальных прав. В случае с фрагментом (2) обоснования модели и суда апелляционной инстанции тоже расходятся, но в этом случае оценить правомерность версии модели можно только с помощью дополнительной экспертизы. Соответственно, при совпадении пункта статьи, предложенного моделью, с «эталонным» пунктом, её ответ на вопрос 2 засчитывался как правильный, а расхождения в обоснованиях не учитывались.

В некоторых случаях модели допускали явные ошибки. БЯМ не всегда выдают ответ в формате, который предписывался им в запросе (например, просто не приводят номер пункта статьи, как это требовалось в промпте). Ответ модели в отдельных случаях является внутренне противоречивым. Например, модель ответила «да» на вопрос 1 (существуют ли основания для пересмотра дела?), но при этом наиболее вероятным решением апелляционной инстанции назвала «оставить решение суда первой инстанции в силе».

Итоговые результаты по количеству «правильных» ответов в каждой категории можно увидеть в таблице 3. Для каждой модели приводится результат по вопросам 1 и 2 промпта («*есть ли основания для пересмотра дела...*» и «*каким должно быть решение суда апелляционной инстанции...*»). Выделены ячейки с максимальным количеством «правильных» ответов в каждой категории.

Таблица 3. Количество «правильных» ответов в тесте 1 по категориям, шт.

Модель	Grok		DeepSeek		ChatGPT		GigaChat		YandexGPT		дел
Вопрос	1	2	1	2	1	2	1	2	1	2	
Административные	5	3	4	4	**6**	**5**	5	4	5	3	10
Гражданские	7	**8**	6	6	8	**8**	**9**	**8**	5	5	13
Уголовные	5	1	7	**3**	**9**	2	4	2	4	2	9
Всего	17	12	17	13	**23**	**15**	18	14	14	10	32

В таблице 4 приводится общий результат для каждой модели в процентах.

Таблица 4. Точность ответов моделей на вопросы теста 1

	Grok	DeepSeek	ChatGPT	GigaChat	YandexGPT
Вопрос 1	56%	53%	**72%**	56%	44%
Вопрос 2	31%	40%	**47%**	44%	31%

В отличие от вопроса 1, для которого предусматривалось только два варианта ответа, на вопрос 2 модель должна была точно назвать номер пункта статьи, в соответствии с которым должно быть вынесено решение апелляционным судом. Результаты по этому вопросу у всех моделей оказались неудовлетворительными, «неверных» ответов больше, чем «верных». Подробные результаты по каждому делу см. в таблице 4 *на онлайн-ресурсе, ссылка на который размещена во введении*.

2.2. Проверка последовательности работы моделей

Второй эксперимент на представленном датасете – тест 2 – заключался в том, чтобы проверить, насколько БЯМ последовательны в своих рассуждениях и выводах. В промпте для второго теста модели были даны как решение суда первой инстанции, так и апелляционное определение по этому же делу.

Модели предлагалось ответить на вопрос, согласна ли она с представленным определением апелляционной инстанции (напомним, выводы апелляционной инстанции мы считали «эталоном» в тесте 1). Если же модель не согласна с предъявленным определением, то в промпте содержалась просьба составить альтернативный проект апелляционного определения, а также указать несколько аргументов, почему модель не согласна с приложенным к промпту апелляционным определением.

Как можно понять из описания эксперимента, целью таких запросов было проверить, насколько на решения БЯМ влияют выданные в промпте в явном виде апелляционные определения, вынесенные судами, меняет ли модель (и насколько часто) своё мнение по сравнению с результатами теста 1, где апелляционное определение моделям не предъявлялось.

В промпте теста 2 моделям было задано два вопроса:

(*1*) после предъявления «эталона» какое мнение у модели по вопросу, есть ли основания для изменения или отмены решения апелляционной инстанцией (да/нет)?

(*2*) после предъявления «эталона» по какому пункту модель предлагает вынести решение апелляционного суда?

Результаты теста 2 представлены в таблице 5.

Таблица 5. Результаты теста 2.

Модель	Grok		DeepSeek		ChatGPT		GigaChat		YandexGPT	
Вопрос	1	2	1	2	1	2	1	2	1	2
Ответ модели совпал с ответом в тесте 1:										
	16	14	20	18	21	17	13	9	14	12
Как изменилось число «правильных» ответов:										
	+4	+16	-6	-6	+1	+6	+7	+7	+2	+3

Нужно прежде всего отметить, что ответы ни одной БЯМ на промпт в тесте 2 не совпали с ответами в тесте 1, то есть предъявленное апелляционное определение повлияло на ход рассуждений и ответ моделей. При этом и ни одна БЯМ не согласилась с эталоном для всех предъявленных в промптах документов. Однако есть и случаи, когда ответ БЯМ не совпадал ни с выводами апелляционного определения, вынесенного судом, ни с выводами самой БЯМ по итогам теста 1.

В процессе тестирования возникали случаи, когда БЯМ ссылалась на несуществующий пункт соответствующего кодекса. Например, модель предлагала отправить дело на пересмотр в суд первой инстанции, но при этом в кодексе в статье о полномочиях апелляционной инстанции такой пункт отсутствует (так в статье 328 ГПК РФ).

На исследованном материале датасета судебных дел дел оказалось, что больше всего подвержена влиянию предъявленного «эталона» модель Grok, менее всего – ChatGPT и DeepSeek. При этом ChatGPT наиболее последователен в своих решениях (если сравнивать результаты тестов 1 и 2), а DeepSeek склонен принимать решения, отличные как от «эталона», так и от результатов своих же рассуждений в тесте 1.

2.3. Оценка решений нескольких видов инстанций

Часто рассмотрение судебных дел происходит последовательно сразу в нескольких видах инстанций – в суде первой инстанции, апелляционном, кассационном и др., вплоть до Верховного суда РФ (ВС РФ). Целью теста 3 было установить, насколько большие модели способны оценивать последовательность решений нескольких видов инстанций по одному и тому же делу.

Данный тест выполнялся со следующим ограничением: поскольку часто не все решения судов по одному и тому же делу находятся в открытом доступе, то данный тест проводился с использованием изложений решений нескольких судов, которые имеются на *сайте Российского агентства правовой и судебной информации*[11] в разделе «Судебная практика». В качестве эталонного ответа бралось решение Верховного суда по этому делу, которое также имеется в указанном выше разделе. Решение апелляционной комиссии (если таковое было) к решению ВС РФ в данном тесте не рассматривалось. Полные тексты изложений дел, которые использовались в этом эксперименте, а также ответы моделей, можно найти *на онлайн-ресурсе, описанном во введении*.

Всего для данного теста было выбрано 10 дел всех трёх категорий – гражданской, административной и уголовной. Они относились к следующим темам:

(1) Условия прекращения краткосрочных договоров аренды.
(2) Отказ от легализации здания из-за долгостроя.
(3) Защита прав жертвы кибермошенников.
(4) Покупка франшизы без прав на ноу-хау и товарный знак.
(5) Конфискация переведённых на счёт членов семьи коррупционера денег.

[11] www.rPSInews.ru

(6) Возможность открытия вклада на имя своего знакомого без его согласия.
(7) Уточнение разницы между взяткой и откатом.
(8) Сохранение прав покупателя на дом при возврате продавцом оплаты.
(9) О толковании в пользу потребителя сомнений в договоре страхования.
(10) Уточнение, в каких случаях замощённую площадку можно считать объектом недвижимости.

Промпт для выполнения теста выглядел следующим образом:

I. Сегодня ты юрист, который будет анализировать решения нескольких судов. Тебе будет дано изложение сути дела и решения нескольких инстанций, которые его рассматривали.

II. Твоя задача:

1) проанализировать данное ниже изложение дела,

2) вынести по нему окончательное решение с его обоснованием.

Само изложение рассматриваемого дела приводилось после текста промпта. Фактическая информация о решении ВС, имеющаяся в изложении дела типа «Определение ВС РФ по делу № 127–КГ25–8–К4», из него убиралась.

Ограничений на формат ответов моделям не задавалось, поэтому их ответы на такое тестовое задание обычно включали в себя несколько разделов:

- Фактические обстоятельства дела.
- Правовая оценка обстоятельств (применение норм соответствующего законодательства).
- Вывод.
- Окончательное решение.

Ответ модели брался из раздела 4 «Окончательное решение» и вручную сравнивался с решением Верховного суда РФ. Шкала оценок совпадения ответа модели с решением Верховного суда была четырёхзначной:

0 – отказ от ответа,
1 – полное совпадение,
2 – частичное совпадение,
3 – несовпадение.

Отказ от ответа был получен только в одном случае – модель запросила дополнительные материалы по делу: «Для точного ответа на вопрос требуется более детальный анализ материалов дела, включая судебные акты и техническую документацию.»

Таблица 6. Результаты Теста 3

Дело	Тема	Grok-4	DeepSeek	GPT-5	GigaChat	YandexGPT
ВС 1	условия прекращения договоров аренды	3	3	3	3	3
ВС 2	отказ от легализации здания из-за долгостроя	1	3	3	1	3
ВС 3	защита прав жертвы кибермошенников	1	1	3	3	1
ВС 4	франшиза bubble tea	2	2	3	3	2
ВС 5	конфискация переведённых на счёт членов семьи коррупционера денег	3	2	2	2	2
ВС 6	для открытия вклада на имя знакомого не нужно его согласие	2	3	3	3	3
ВС 7	уточнение разницы между взяткой и «откатом»	2	3	3	2	3
ВС 8	возврат продавцом оплаты не лишает покупателя прав на дом	1	1	1	1	3
ВС 9	сомнения в договоре страхования должны толковаться в пользу потребителя	2	2	3	3	3
ВС 10	уточнение, в каких случаях замощённую площадку можно считать объектом недвижимости	1	3	3	1	0

Из представленных результатов можно видеть, что точность моделей на решении задачи оценки решений нескольких видов инстанций по одному и тому же делу оказалась очень невысокой, что свидетельствует о том, что современные БЯМ еще не готовы для применения в реальной судебной практике. Если сравнивать модели между собой, то лучшей на данном тесте оказалась система Grok, далее следует GigaChat, потом DeepSeek, и в конце примерно на одном уровне GPT–5 и YandexGPT.

2.4. Квалификация преступных деяний

2.4.1. *Вводные замечания*

Цель теста – оценить способность моделей давать некоторому деянию уголовно-правовую квалификацию.

Тест заключается в следующем. На вход модели подаётся фрагмент описательно-мотивировочной части обвинительного приговора суда, излагающий фактические обстоятельства дела, но не содержащий подсказок относительно правовой квалификации деяния. Модель должна установить, в совершении каких преступлений, предусмотренных Уголовным кодексом РФ, обвиняется подсудимый. Предсказанные моделью правовые нормы сравниваются с приведённым в резолютивной части приговора перечнем статей, по которым обвиняемого признал виновным суд.

2.4.2. *Исходные данные для теста*

Исходными данными для теста являются 10 документов, содержащих описания преступных деяний, и соответствующие им эталонные множества квалификаций, которые использовались для проверки ответов моделей (подробнее об этих квалификациях будет сказано далее). Данные для теста сформированы на основе реальных приговоров по уголовным делам, содержащих обвинения по различным статьям УК РФ. Их отбор осуществлялся согласно следующим критериям:

Критерий 1. Отбирались только обвинительные приговоры, поскольку в результате теста модель должна предсказать конкретные нормы Уголовного кодекса РФ.

Критерий 2. Рассматривались приговоры только с одним обвиняемым.

Критерий 3. Правовые нормы, приведённые в резолютивной части приговора, соответствуют действующей редакции УК РФ и не претерпевали существенных изменений с момента вынесения приговора.

Описание преступного деяния представляет собой фрагмент описательно-мотивировочной части приговора, который

- содержит описание фактических обстоятельств дела, влияющих на квалификацию деяния, но

- не содержит явных указаний на уголовно-правовую квалификацию содеянного: является ли это деяние преступлением и в чем именно суд усматривает состав преступления.

Все подобные формулировки удалялись вручную из текста приговора и заменялись обозначением «[. . .]».

Соответственно, во фрагменте сохранялись указания на место, время и способ совершения преступления. Сохранялись также описания, указывающие на различные элементы состава преступления (форма вины, тяжесть последствий, цели, мотивы), но без формулировок с юридической оценкой последних.

Детали, не связанные с составом преступления и не влияющие на квалификацию (личные данные подсудимого, информация о мере пресечения, анализ доказательств и их оценка, смягчающие и отягчающие обстоятельства, процессуальные аспекты и др.) из фрагмента исключались.

Предполагается, что полученное таким образом описание содержит информацию, достаточную для установления того, является ли описанное деяние преступлением и какими пунктом, частью, статьей Уголовного кодекса Российской Федерации оно предусмотрено.

2.4.3. *Промпт*

Модели предлагался следующий промпт:

> *Сегодня ты юрист. Тебе дан фрагмент текста приговора с купюрами. На основании этого фрагмента ответь: в совершении каких преступлений, предусмотренных УК РФ, признан виновным подсудимый?*

Промпт не содержал никаких требований, ограничивающих свободу модели в формулировании ответа. Как следствие, в ответах моделей, наряду с предсказанными правовыми нормами, присутствовала также и дополнительная информация, полезная для содержательного анализа ответов, – правовые описания, ход рассуждений, обоснование выбора того или иного решения.

2.4.4. *Множества предсказанных и эталонных квалификаций*

Для удобства сравнения уголовно-правовых норм, предсказанных моделью, с эталонными (содержащимися в приговоре) и те, и другие представляются в виде множества *единиц квалификации*. В качестве основных единиц квалификации могут выступать:

- номер статьи (если не указаны номер части статьи и буквенные обозначения пунктов);
- номер статьи и её части (если не указаны пункты);

- номер статьи, номер части и пункт.

В случае если состав преступления включает несколько квалифицирующих признаков, что соответствует обвинению по нескольким пунктам одной части статьи УК (или по нескольким частям одной статьи), за единицу квалификации принимается отдельная комбинация вида «пункт-часть-статья» (или «часть-статья»). Например, преступлению, предусмотренному **пп. «а», «г» ч. 4, ч. 5 ст. 228.1 УК РФ**, соответствуют три единицы квалификации:

- **п. «а» ч. 4 ст. 228.1;**
- **п. «г» ч. 4 ст. 228.1;**
- **ч. 5 ст. 228.1.**

В особых случаях, когда в квалификации деяния присутствует некоторая норма, которая никогда не применяется сама по себе[12] (например, **часть 1** или **часть 3 ст. 30** – приготовление к совершению преступления или покушение на совершение преступления), то она добавляется к каждой из единиц квалификации, составляющих связанную с ней норму – конкретный состав преступления. Так, квалификация преступления по **ч. 3 ст. 30, пп. «а», «в», «е» ч. 2 ст. 105 УК РФ** будет разделена на следующие единицы:

- **ч. 3 ст. 30 + п. «а» ч. 2 ст. 105;**
- **ч. 3 ст. 30 + п. «в» ч. 2 ст. 105;**
- **ч. 3 ст. 30 + п. «е» ч. 2 ст. 105.**

Далее для краткости единицы квалификации будем называть просто *квалификациями*.

Множество предсказанных квалификаций формировалось следующим образом. Из текста, сгенерированного моделью в ответ на вопрос промпта, извлекались правовые нормы, которыми она сочла нужным квалифицировать преступление, и представлялись в виде множества (единичных) квалификаций. В результат засчитывались только итоговые квалификации. Квалификации, упоминавшиеся моделью в ходе рассуждений, но не подтверждённые в заключительной части ответа, игнорировались.

2.4.5. *Методика оценки результатов*

Оценивать результат предсказания с помощью количественных метрик типа F1-мер мы сочли нецелесообразным, поскольку это означало бы свести все несовпадения с эталоном к ложным срабатываниям (false positives) и пропущенным целям (false negatives). Даже специальная

[12] Речь идёт о положениях Общей части УК РФ, которые применяются только с какой-либо нормой Особенной части.

hF1-мера для иерархически организованных классов, позволяющая учитывать частичное совпадение с точностью до некоторого уровня, остаётся всё же только количественной мерой и не даёт возможности отразить разнообразие предсказанных моделями результатов.

В основе используемой нами классификации несовпадений лежит различение случаев соотносимости и несоотносимости предсказания и эталона. В группе соотносимых случаев выделяется три вида ошибок: огрубление, чрезмерная детализация, расхождение (таблица 7), в группе несоотносимых случаев – два вида ошибок: лишняя квалификация и квалификация отсутствует (таблица 8). Дифференциация ошибок внутри каждого вида продиктована следующим соображением: поскольку Уголовный кодекс имеет иерархическую структуру, вполне естественным кажется предположение, что ошибка тем значимей, чем

(*1*) выше уровень ошибки,
(*2*) больше разница в уровне между предсказанием и эталоном[13].

При оценке результатов теста для каждой эталонной квалификации делается попытка найти совпадающую или иную совместимую с ней предсказанную квалификацию. Лишней предсказанная квалификация признается, если для всех эталонных квалификаций найдены бесспорно сопоставимые предсказания. Отсутствие предсказания для некоторой эталонной квалификации усматривается, если все предсказания оказались бесспорно сопоставимы с другими эталонами[14].

2.4.6. *Анализ результатов теста*

Недостатком детальной классификации ошибок для теста приходится признать то, что она не даёт возможности ранжировать все модели по некоторому единому интегральному показателю. Однако, учитывая скромный масштаб эксперимента, в этом нет особой необходимости.

Заметим также, что эта классификация носит формальный характер: сравниваются исключительно идентификаторы уголовно-правовых норм – номера статей, их частей и буквенные индексы пунктов. О содержательной оценке ответов моделей мы поговорим отдельно.

Результаты сравнения предсказанных квалификаций с эталонными для пяти моделей и 10 дел сведены в таблицу 9. Суммарное число эталонных квалификаций – 16.

---

[13] Расстояние между уровнями релевантно только для ошибок вида «огрубление».
[14] Не пытаясь строго определить, что значит «бесспорно сопоставимы», заметим, что сложностей с установлением лишних и недостающих квалификацией не возникало.

Таблица 7. Ошибки при несовпадении эталонной и предсказанной квалификаций

Описание ситуации		Обозначение ошибки
Огрубление:		
	верно предсказаны статья и часть, но пункт не указан	часть вместо пункта
	верно предсказана статья, но часть не указана	статья вместо части
	верно предсказана статья, но часть и пункт не указаны	статья вместо пункта
Чрезмерная детализация:		
	верно предсказаны статья и часть, но указан еще и пункт, которого в эталоне нет	пункт вместо части
	верно предсказана статья, но указана еще и часть, которой в эталоне нет	часть вместо статьи
Расхождение:		
	верно предсказаны статья и часть статьи, но пункт отличается от эталона	другой пункт
	верно предсказана статья, но часть отличается от эталона	другая часть
	предсказана другая статья	другая статья

Таблица 8. Ситуации несопоставимости эталонной и предсказанной квалификаций

Описание ситуации	Обозначение ошибки
Квалификация отсутствует	
Случай 1: Не предсказан один из двух или более пунктов части статьи и нет сопоставимого с ним	нет пункта
Случай 2: Не предсказана одна из двух или более частей статьи и нет сопоставимой	нет части
Случай 3: Для данной эталонной квалификации нет сопоставимого предсказания и это не случай 1 или 2	нет статьи
Лишняя квалификация	
Случай 1: Предсказанный пункт части статьи не сопоставим ни с одним эталонным (при том что другие пункты той же части сопоставимы)	лишний пункт
Случай 2: Предсказанная часть статьи не сопоставима ни с одной эталонной (при том что другие части той же статьи сопоставимы)	лишняя часть
Случай 3: Предсказанная квалификация не сопоставима ни с одной эталонной квалификацией и это не случай 1 или 2	лишняя статья

Таблица 9. Результаты сравнения ответов БЯМ с эталонными квалификациями

Вид ошибки	Обозначение ошибки	Grok	DeepSeek	ChatGPT 5	GigaChat	YandexGPT
Полное совпадение						
	ошибок нет	11	9	7	2	9
Огрубление						
	часть вместо пункта	1	1	1	1	2
	статья вместо части		2	3	4	4
	статья вместо пункта		1		2	1
Чрезмерная детализация						
	часть вместо статьи				1	
Расхождение						
	другой пункт		1	1	1	
	другая часть	1		2	1	
	другая статья	3	2	2	3	
Квалификация отсутствует						
	нет статьи				1	
Лишняя квалификация						
	лишний пункт					1
	лишняя часть			1		
	лишняя статья	1	1	1	4	5

По числу *правильно предсказанных квалификаций* лидирует Grok. DeepSeek, YandexGPT и ChatGPT делят второе место, а GigaChat показал очень скромный результат.

Ошибки вида «огрубление». И здесь Grok оказался лучше всех, допустив всего одно минимальное огрубление. На втором месте ChatGPT и DeepSeek (4 случая разной тяжести). У YandexGPT результат заметно хуже (7 ошибок этого вида), еще хуже у GigaChat (тоже 7 ошибок, но из них два случая, когда не предсказаны ни пункт, ни даже часть эталонной статьи).

Расхождения с эталонной квалификацией. YandexGPT не допустил ни одного случая расхождения. У всех остальных моделей такие случаи имели место, от различий в пункте правильно предсказанной части статьи до предсказания альтернативной статьи.

Меньше всего *лишних квалификаций* предсказали Grok и DeepSeek. Им немного уступает ChatGPT. В 4 случаях лишние статьи предсказал GigaChat, а худший результат у YandexGPT (один лишний пункт в дополнение к правильно предсказанному и 5 лишних полноценных квалификаций).

GigaChat оказался единственным, кто допустил *ошибку чрезмерной детализации* (указал часть статьи, когда нужна была просто статья) и *ошибку вида отсутствие квалификации* (не предсказал одну из двух эталонных статей).

Таким образом, по результатам *формального сравнения* предсказанных моделями квалификаций с эталонными, по совокупности показателей лидером следует признать Grok. Хуже других с заданием теста справился GigaChat.

2.4.7. *Содержательный анализ результатов*

Если обратиться к содержанию норм закона (к диспозициям соответствующих статей УК), то большинство случаев несовпадения предсказаний с эталонами выглядят как неполные (недостаточно конкретные) или альтернативные[15] квалификации. Иными словами, предсказания моделей по бóльшей части носят вполне осмысленный характер.

К примеру, предсказание может оказаться квалификацией по основному составу преступления вместо квалифицированного состава, как того требует эталон, либо содержит не все квалифицирующие признаки.

Например, модель предсказала **ч. 1 ст. 272** – Неправомерный доступ к охраняемой законом компьютерной информации, если это деяние повлекло

[15] Мы не обсуждаем здесь вопрос об их юридической правомерности.

уничтожение, блокирование, модификацию либо копирование компьютерной информации [...], тогда как суд квалифицировал преступление по **ч. 2 ст. 272** – То же деяние, причинившее крупный ущерб или совершенное из корыстной заинтересованности.

Модель может предложить квалификацию по статье, содержательно близкой к эталонной (напр., вместо **ст. 322.3** – *Фиктивная постановка на учет иностранного гражданина или лица без гражданства по месту пребывания в Российской Федерации* предсказана **ст. 322.2** – *Фиктивная регистрация гражданина Российской Федерации по месту пребывания или по месту жительства в жилом помещении в Российской Федерации и фиктивная регистрация иностранного гражданина или лица без гражданства по месту жительства в жилом помещении в Российской Федерации).*

Расхождение в выборе пункта статьи может быть объяснено конкуренцией квалифицированных составов: так, убийство путем обливания легковоспламеняющейся жидкостью и поджога модель квалифицировала по **п. «е» ч. 2 ст. 105** – *Убийство, совершенное общеопасным способом*, вместо эталонного **п. «д»** – *Убийство, совершенное с особой жестокостью.*

Отличие предсказанной статьи от эталонной может быть представлено как выбор моделью общей нормы при конкуренции общей и специальной норм, напр. квалификация по **п. «б» ч. 2 ст. 105** – *Убийство лица или его близких в связи с осуществлением данным лицом служебной деятельности или выполнением общественного долга* вместо эталонной **ст. 317** – *Посягательство на жизнь сотрудника правоохранительного органа.*

Содержательных ошибок-галлюцинаций в результатах теста отмечено немного. Ошибка чрезмерной детализации, допущенная GigaChat, оказалась квалификацией по несуществующей части реальной статьи (в **ст. 322.3** нет части 1). Признаки галлюцинаций обнаружены в двух случаях расхождения с эталоном в результатах ChatGPT, а также в лишнем пункте у YandexGPT: предсказанные квалификации содержательно явно не соответствуют эталонным.

2.4.8. *За рамками теста*

Как уже говорилось, промпт не накладывал никаких ограничений на формат ответов моделей. Анализ полнотекстовых версий ответов позволил увидеть то, что не нашло отражения в результатах теста.

Главное наблюдение состоит в том, что модели часто допускают ошибку *идентификации нормы*: выдаваемый в качестве квалификации идентификатор нормы по УК не соответствует вербальному описанию

квалифицирующего состава, приведенному в ответе. Это особенно проявляется на уровне идентификации квалифицирующих признаков (выбор пункта части статьи, реже – части статьи).

Модель может совершенно правильно квалифицировать факты на уровне формулировок, но в результате выдать не соответствующий им идентификатор нормы. Так, ответ DeepSeek по одному из дел расходится с эталоном (предсказан **пункт «е» части 2 ст. 105** вместо **пункта «д»**), при том что и в обосновании, и в заключении модель упоминает квалифицирующий признак «с особой жестокостью», в точности соответствующий эталонному **пункту «д»**. ChatGPT в ответе по этому же делу убедительно, ссылаясь на факты, обосновывает тот же квалифицирующий признак «с особой жестокостью», но выдает его за **пункт «а»** (что в УК соответствует убийству «двух и более лиц»).

Возможно и другое проявление той же ошибки. К примеру, DeepSeek и Grok словесно описывают признак, в точности соответствующий эталонному пункту квалификации, но «забывают» обозначить его буквой в итоговой квалификации.

А в одном случае ошибка идентификации даже привела к правильному ответу. YandexGPT усмотрел более мягкий состав преступления по сравнению с эталоном, но «перепутал» номера статей: её формулировке (*«применение насилия в отношении представителей власти»*) соответствует **ст. 318**, но модель назвала это **ст. 317**, что и оказалось правильной квалификацией (**ст. 317** – *Посягательство на жизнь сотрудника правоохранительного органа*).

2.4.9. *Вывод*

Таким образом, в тесте на квалификацию преступных деяний большие языковые модели продемонстрировали как свои сильные стороны (способность рассуждать, опираясь не только на изложенные в деле факты, но и на общие представления о мире, сопоставлять факты и собственные выводы с описаниями норм закона), так и свои слабости. Самым серьезным недостатком БЯМ при выполнении тестового задания следует признать ненадежность идентификации моделью реальной уголовно-правовой нормы, соответствующей сгенерированному текстовому описанию.

3. Смежные работы

Тема прогнозирования судебных решений давно присутствует как в исследовательской области, так и на практике. Так, в работе [16] приводится систематический обзор международных исследований, посвящённых возможности использования искусственного интеллекта для прогнозирования

решений судов. Автор рассматривает выборку из 107 международных публикаций за 2004—2023 гг. на тему классификации судебных текстов; из них 34 источника за 2015–2023 гг. анализирует количественно, с точки зрения динамики качества предсказательных моделей. Подавляющее большинство моделей решает задачу предсказания в «слабом» смысле: алгоритм обучается и тестируется на разных подмножествах одного и того же набора данных (т. е. на уже имеющихся решениях).

Отмечается, что с развитием технологий ИИ и появлением открытых баз судебных актов точность прогнозов заметно выросла. Так, показатель для китайского датасета CAIL2018 вырос с 78% (2018 г.) до 96% (2023 г.), а для решений Верховного суда США – с 70% (2015—2017 гг.) до 92,5% (2022). При этом, как правило, оценивается точность (accuracy) предсказания бинарного исхода: виновен / не виновен, жалоба удовлетворена / отклонена, апелляция успешна / нет и т. д.

Что касается методов, то самыми актуальными на момент публикации статьи А. П. Казуна [16] признавались предобученные трансформеры типа BERT и построенные на их основе специализированные модели, дообученные на юридических текстах (LegalBERT). Их главный недостаток – невозможность интерпретировать, на основании каких факторов был сделан прогноз, что ставит принципиальный барьер на пути к возможной замене судьи искусственным интеллектом. Кроме того, модели имеют низкую внешнюю валидность, а это значит, что алгоритмы, хорошо предсказывающие вердикты Верховных судов, не могут быть использованы для массовых дел нижестоящих инстанций.

Это обстоятельство, вкупе с дефицитом качественных данных по типовым категориям дел, приводит к тому, что модели оказываются неприменимы там, где они нужнее всего. Дополнительные опасения связаны с вопросами этики и доверия: обучаясь на прецедентах, модели воспроизводят не только нормы, но и «психологию судебного разбирательства», включая расовые, гендерные и иные стереотипы. Автор приходит к выводу, что в обозримом будущем полная замена судьи-человека ИИ маловероятна, однако ИИ–технологии могут стать эффективным вспомогательным инструментом, снижающим рутинную нагрузку и высвобождающим время для содержательного анализа дела.

Однако, поскольку в начале 2025 г. произошел коренной сдвиг в разработке больших языковых моделей и появились модели с механизмом рассуждений, то большинство работ, которые опирались в основном на BERT–системы, утратили свою актуальность и полезность. Подавляющее большинство современных работ базируется на использовании БЯМ. Если традиционно БЯМ использовались для промежуточных целей, таких как извлечение фактов или разметка датасета [21, 22], то сейчас наметилось

несколько направлений: тестирование возможностей БЯМ; разработка решений для конкретных задач (например, прогнозирование решения суда); создание датасетов для оценки способности БЯМ к рассуждению в отдельных областях.

Нередко в подобных экспериментах используются модифицированные данные, как, например, в случае [**15**]. В этой работе использовалось реальное судебное дело, в котором был изменен обвиняемый (в одной модификации он должен был «вызывать сочувствие», в другой «не заслуживал сочувствия») и прецеденты, чтобы определить, что именно влияет на принятие решений судьями. Оба варианта были предложены GPT–4o, которая должна была определить, виновен подсудимый или нет.

В результате оказалось, что GPT–4o больше ориентируется на прецедент, тогда как реальные судьи в аналогичном эксперименте в основном ориентировались на личность подсудимого. Основным недостатком этой работы, который отмечают сами авторы, является то, что рассматривалось лишь одно дело и одна БЯМ, что не позволяет делать общий вывод о способности БЯМ к вынесению судебных решений.

Также для улучшения результатов предсказания судебного решения строятся многоступенчатые системы обработки данных. Одна из таких систем – LegalReasoner – представлена в статье [**23**]. Предлагается многоэтапный пайплайн на основе БЯМ Llama2. Результат его работы авторы сравнили с результатами аналогичных систем, в основе которых лежали нейросети предыдущих поколений, без рассуждений, а также с результатами работы специализированной БЯМ Legal–PEGASUS, обученной на юридических данных.

Цель работы является практической – повысить процент правильных предсказаний по сравнению с конкурентами. Для этого выстроен многоступенчатый алгоритм: дообученные на юридических текстах нейросети, использование контрастного, генеративно-состязательного и многозадачного обучения для разных этапов пайплайна (соответственно, для разных нейросетей).

Проведенные эксперименты показали, что результаты LegalReasoner превосходят результаты аналогичных систем примерно на 8%. Нужно отметить, что на вход системе LegalReasoner должен поступать обработанный текст – краткое фактическое описание дела (для обучения и тестирования работы LegalReasoner были взяты датасеты, в которых такие описания уже подготовлены). На выходе система предсказывает исход дела – виновен/не виновен, статью, по которой выносится обвинение, и срок заключения, если таковой имеется, качество рассуждений никак не оценивается.

В статье [24] рассматривалось «применение больших языковых моделей для юридических экспертиз». В ней предпринята попытка продолжить работу по оценке способностей БЯМ к логическим рассуждениям [25]. Рассматривались «юридические кейсы», связанные с правовыми аспектами разработки и использования программного обеспечения (ПО).

Отбирались случаи, иллюстрирующие некоторые парадоксы в деонтической логике, в которой формализуются рассуждения с использованием понятий долженствования, обязательств и запрещения. Все отобранные случаи, как утверждают авторы, «приводят к противоречию или интуитивно непостижимой ситуации», однако это противоречие лежит не строго в логической или нормативно-правовой плоскости, а возникает при столкновении ситуации со здравым смыслом или моралью.

Для каждого случая были подготовлены словесное описание, иногда включавшее в себя фрагменты нормативно-правовых документов (договоров, контрактов, должностных инструкций и т. п.); формулировка проблемы, вытекающей из описанной ситуации; описание с помощью формул деонтической логики, но без объяснения привязки имен пропозиций (атомов) из этих формул к утверждениям из описания кейса. Тестируемые модели получали на вход описание кейса и должны были спрогнозировать решение юриста по заданной проблеме. Ответы моделей оценивались юристами-практиками. Испытывались модели GPT–4o, GigaChat MAX и YandexGPT 3.

Авторы отмечают, что ни одна модель не дала удовлетворительных ответов на заданные вопросы. В частности, было подчеркнуто отсутствие ссылок на нормативно-правовые документы актуального российского законодательства, что в частности объясняется тем, что в исходном промпте не было явных указаний сделать это.

Для устранения отмеченных недостатков авторы предложили:

(*1*) усовершенствованный промпт,

(*2*) скорректированные 6 параметров модели, которые влияют на процесс генерации ею ответа,

(*3*) архитектуру системы, включающую RAG–хранилище нормативно-правовых документов, как обычных (УК, ГК РФ), так и специальных (устав компании, должностные инструкции, и др.).

Анализ решений каких-либо судебных инстанций по рассматриваемым кейсам в данной работе, в отличие от нашей, не проводился.

Другая тенденция в исследованиях по применению БЯМ в юридической сфере – прогнозирование судебных решений [14, 19]. Особо отметим работу [26], где предлагается архитектура, использующая несколько агентов на основе БЯМ. Каждый из этих агентов имеет свою

роль (профессиональный судья, начинающий судья и т. п.) и образуют «судебную коллегию». Получив описание преступления, релевантные нормативные акты и обвинение, они должны «посовещаться» и определить срок тюремного заключения (это одна из подзадач в прогнозировании судебных решений). Эта агентная архитектура может быть использована в дополнение к обычному промпту в БЯМ. Она был протестирована для моделей GPT–3.5, GPT–4, Qwen на существующих датасетах уголовных дел китайских судов. В целом запросы с использованием этой «судебной коллегии» показали немного более высокий результат, чем простые промпты, промпты с примером цепочки рассуждений (chain-of-thought) и др., особенно по параметру «этичность». Однако остается открытым вопрос о том, насколько успешным было бы использование подобной архитектуры в российских юридических реалиях.

Последняя из упомянутых тенденций – разработка датасетов – представлена, в частности, работой [7]. Авторы составили датасет, предназначенный для оценки способностей БЯМ к рассуждению в области налогового права. Датасет был составлен на основе 199 писем Минфина РФ и ФНС, которые содержат подробные ответы на вопросы налогоплательщиков. Датасет имеет следующую структуру: для каждого письма а) был сформулирован вопрос, на который следует ответить «да» или «нет» (ответ следует из содержания письма), эта часть датасета была получена с помощью ChatGPT; б) дан ответ на этот вопрос («да»/«нет»); в) были извлечены ссылки на релевантные нормативные правовые акты (Налоговый кодекс РФ, Федеральные законы, Постановления правительства и т. п.). Уделялось особое внимание тому, чтобы часть вопросов требовали ответа «нет», так как для получения такого ответа моделям требуется более основательное рассуждение.

С помощью этого датасета были протестированы модели Mixtral 8x7B, Llama 3.3 70B и GPT–4o mini. Тестирование проводилось в четырех различных вариантах, где на вход моделям подавался либо вопрос + пример рассуждения, либо вопрос + выдержки из релевантных нормативных актов, либо вопрос + векторное представление релевантных нормативных актов. Модели должны были ответить «да» или «нет» на заданный вопрос. Для самого простого варианта точность ответов моделей составила от 57% до 65%, а для конфигурации с RAG (нормативные акты дословно или в виде векторного представления) точность составила от 66% до 77%.

Статья [27] затрагивает проблему дефицита качественных дезагрегированных данных для эмпирических исследований российского правосудия. В качестве решения предлагается методология использования публикуемых в открытом доступе текстов судебных приговоров как богатого источника данных о правоприменении для социолого-правового анализа.

Авторы детально описывают весь конвейер создания исследовательских массивов данных из первичных текстов: от сбора приговоров с официальных порталов (ГАС «Правосудие», Мосгорсуд) с помощью веб-скрейпинга до их предобработки и автоматизированного извлечения значений переменных. Предлагается комбинировать два подхода: правила и регулярные выражения для стандартизированных формулировок (пол, наличие у обвиняемого малолетних детей, смягчающие и отягчающие обстоятельства) и машинное обучение с учителем для сложных, неформализуемых категорий (например, характер отношений между обвиняемым и жертвой).

В качестве иллюстрации к излагаемым подходам авторы используют исследование гендерных различий в приговорах по **ст. 105 ч. 1 УК РФ** (убийство), в ходе которого из текстов приговоров были извлечены следующие переменные: срок приговора, пол обвиняемого, пол потерпевшего, характер отношений между обвиняемым и потерпевшим (интимные партнеры, родственники, другое), наличие малолетних детей у обвиняемого, степень рецидива, степень признания вины, наличие иных смягчающих и отягчающих обстоятельств. Для анализа были автоматически обработаны тексты 20531 приговора.

Авторы отмечают существенные преимущества текстов приговоров по сравнению с другими доступными источниками данных (такими, как официальная статистика, данные коммерческих и справочно-правовых систем).

Дезагрегированность: в отличие от сводной официальной статистики (отчетов 4–ЕГС, данных Судебного департамента при Верховном суде РФ), приговоры позволяют анализировать отдельное дело, участников дел; выявлять типичные обстоятельства совершения преступлений, выявлять различия между группами преступников, выявлять предвзятость в судебных решениях и т. п.

Содержательная насыщенность: тексты приговоров содержат подробные данные, которые не фиксируются в ведомственной статистике (обстоятельства, мотивы, последствия совершения преступления, показания свидетелей) и которые позволяют извлекать и исследовать в том числе нетривиальные (редко исследуемые) параметры. Кроме того, данные из приговоров можно дополнить информацией из сторонних источников о суде, судье, регионе, что расширяет поле для анализа.

Масштаб и доступность: миллионы документов в открытом доступе обеспечивают высокую статистическую мощность анализа и возможность изучать даже редкие явления (например, детерминанты получения оправдательного приговора или особенности работы судов присяжных).

Среди ограничений упоминаются неизбежные потери данных из-за деперсонализации (часто избыточной); неполнота и несистематичность наполнения (не все подлежащие публикации приговоры выкладываются в открытый доступ); недостоверность (субъективность и возможные искажения в описаниях фактов); низкая информативность текстов дел, рассмотренных в особом порядке. Несмотря на эти ограничения, авторы подчеркивают, что в России пока нет альтернативных открытых источников, совмещающих все перечисленные выше достоинства.

Статья представляет интерес в контексте обсуждения прогнозирования судебных решений, поскольку предлагает конкретные, апробированные методы трансформации «сырых» юридических текстов в структурированные данные, которые впоследствии могут стать основой для построения предсказательных моделей.

4. Выводы и направления дальнейшей работы

В данной работе изложены результаты тестирования больших языковых моделей ChatGPT, Grok, DeepSeek, GigaChat и YandexGPT на задаче анализа и прогнозирования судебных решений. Несмотря но то, что LegalAI уже на текущий момент широко и успешно используется в таких задачах из юридической практики, как

(*1*) анализ обязательственных документов, таких как контракты, договора и т. п., на предмет наличия противоречий, конфликтов интересов и соответствия действующему законодательству,

(*2*) составление предварительных вариантов нормативно-правовых документов (legal document drafting),

(*3*) ответы на вопросы по актуальным законодательным актам, применение его в практической юридической работе, связанной с анализом и предсказанием судебных решений по реальным делам,

является преждевременным. Результаты нашего тестирования показали, что точность ответов больших моделей при решении данной задачи лишь в отдельных случаях превышает 50%.

Все предложенные тесты различаются по методике проведения и оценки, однако в общих чертах возможно подвести итоговые результаты. По результатам всех тестов ранжировать рассмотренные модели можно следующим образом (с учетом сказанного выше о том, что результативность каждой в целом находится на среднем уровне):

(*1*) Grok,

(*2*) GPT-5,

(3) DeepSeek,

(4) Gigachat,

(5) YandexGPT.

Ранжирование проводилось следующим образом. Сначала модели ранжировались по результатам каждого теста, и в зависимости от места начислялись баллы (1 место – 5 баллов, 2 место – 4 балла и т.д.). В том случае, если для нескольких моделей было невозможно распределить места, то высчитывалось среднее арифметическое для двух соседних мест. Например, если две модели показали одинаково высокий результат, максимальный в тесте, то они делят между собой 1 и 2 место, т.е. получают (5+4)/2=4,5 балла. Таким образом, в итоговом ранжировании мы получили минимальный разрыв между местами 1-3 (3,8 балла, 3,5 баллов, 3,3 балла), а также между местами 4 и 5 (2,5 балла и 1,9 баллов).

В качестве направлений дальнейшей работы мы планируем:

1. Исследование возможностей многоагентного подхода к прогнозированию и обоснованию судебных решений, в рамках которого «коллегия судей», состоящая из нескольких больших моделей, каждая со своей собственной ролью, обсуждает, анализирует юридический кейс и принимает по нему совместное решение;
2. Исследование возможностей применения RAG-систем, заставляющих большие модели принимать свои решения на основе заданных, фиксированных множеств актуальных нормативно-правовых документов.

Список использованных источников

[1] Цшайге Х. *Автоматизация юридических департаментов: итоги опроса LegalInsight* // LegalInsight.– 2023.– № 10 (126).– С. 18–25. URL URL ↑22

[2] *Опыт юрдепов, которые уже применяют ИИ: реальные кейсы и задачи.*– 2025. URL ↑22

[3] Couture R. J. *The impact of artificial intelligence on law firms' business models*, February 25.– Center on the Legal Profession, Harvard Law School.– 2025. URL ↑22

[4] White S. *How AI is reshaping the future of legal practice.*– The Law Society.– 2024. URL ↑22

[5] *Legal innovation: Seizing the future or falling behind?* Wolters Kluwer 2024 Future Ready Lawyer Survey Report.– 2024.– 18 pp. URL ↑22

[6] *Gartner identifies the Top 6 use cases for generative AI in legal departments*, February 19.– 2025. URL ↑22

[7] Alibekov A., Matenkov A., Bolshakov V., Mukhtarova G., Migal A., Muryshev A., Kozachenko A., Mikhaylovskiy N. *RuTaR — A dataset in Russian for reasoning about taxes // Proceedings of the International Conference «Dialogue 2025»* (Moscow, Russia, 23–25 April 2025), Computational Linguistics and Intellectual Technologies.– vol. **23**.– 2025.– Pp. 12–27. ↑23, 47

[8] Горбунов О. Н. *От паттернов к прецедентам – применение больших языковых моделей в юридической практике, технические и практические аспекты* // Судья.– 2026.– № 1. ↑23

[9] Fan Y., Ni J., Merane J., Tian Y., Hermstrüwer Y., Huang Y., Akhtar M., Salimbeni E., Geering F., Dreyer O., Brunner D., Leippold M., Sachan M., Stremitzer A., Engel Ch., Ash E., Niklaus J. *LEXam: benchmarking legal reasoning on 340 law exams.*– 2025.– 38 pp. arXiv 2505.12864 ↑23

[10] Shen J., Xu J., Hu H., Lin L., Zheng F., Ma G., Meng F., Zhou J., Han W. *A law reasoning benchmark for LLM with tree-organized structures including factum probandum, evidence and experiences // Findings of the Association for Computational Linguistics: ACL 2025* (Vienna, Austria, July 27 — August 1 2025).– ACL.– 2025.– ISBN 9798331323936.– Pp. 17252–17274. arXiv 2503.00841 ↑23

[11] Medvedeva M., McBride P. *Legal judgment prediction: if you are going to do it, do it right // Proceedings of the Natural Legal Language Processing Workshop* (Singapore, 7 December 2023).– ACL.– 2023.– ISBN 9781713886044.– Pp. 73–84. ↑23

[12] Sesodia M., Petrova A., Armour J., Lukasiewicz T., Camburu O.-M., Dokania P., Torr P., Schröder de Witt C. *AnnoCaseLaw: a richly-annotated dataset for benchmarking explainable legal judgment prediction.*– 2025.– 15 pp. arXiv 2503.00128 ↑23

[13] Nigam S. K., Balaramamahanthi D. P., Mishra S., Shallum N., Ghosh K., Bhattacharya A. *NyayaAnumana and INLegalLlama: the largest Indian legal judgment prediction dataset and specialized language model for enhanced decision analysis // Proceedings of the 31st International Conference on Computational Linguistics* (Abu Dhabi, United Arab Emirates, 19–24 January 2025).– ACL.– 2025.– ISBN 979-8-3313-1356-2.– Pp. 11135–11160. ↑23

[14] Kmainasi M. B., Shahroor A. E., Al-Ghraibah A. *Can large language models predict the outcome of judicial decisions?*– 2025.– 6 pp. arXiv 2501.09768 ↑23, 46

[15] Posner E. A., Saran S. *Judge AI: assessing large language models in judicial decision-making*, Coase-Sandor Institute for Law & Economics Research Paper Series, No 25-03.– Coase-Sandor Institute for Law & Economics.– 2025.– 41 pp. ↑23, 45

[16] Казун А. П. *Может ли искусственный интеллект прогнозировать решения суда? Систематический обзор международных исследований* // Мониторинг общественного мнения: экономические и социальные перемены.– 2024.– № 5.– С. 100–122. ↑23, 43, 44

[17] Мазуков М. *Нейросети оценили дело Долиной.*– 2025. ↑23

[18] Мазуков М. *Нейросеть для юристов. Как ИИ помог отменить решение суда первой инстанции.*– 2025. ↑23

[19] Nigam S. K., Patnaik B. D., Mishra S., Thomas A. V., Shallum N., Ghosh K., Bhattacharya A. *NyayaRAG: realistic Legal Judgment Prediction with RAG under the Indian common law system.*– 2025.– 18 pp. arXiv 2508.00709 ↑25, 46

[20] Laban P., Hayashi H., Zhou Y., Neville J. *LLMs get lost in multi-turn conversation.*– 2025.– 36 pp. arXiv 2505.06120 ↑25

[21] Min H., Noh B. *TRACS-LLM: LLM-based traffic accident criminal sentencing prediction focusing on imprisonment, probation, and fines* // Artificial Intelligence and Law.– 2025.– 22 pp. ↑44

[22] Wei B., Yu Y., Gan L., Wu F. *An LLMs-based neuro-symbolic legal judgment prediction framework for civil cases* // Artificial Intelligence and Law.– 2025.– 35 pp. ↑44

[23] Wang X., Zhang X., Hoo V., Shao Z., Zhang X. *LegalReasoner: A multi-stage framework for legal judgment prediction via large language models and knowledge integration* // IEEE Access.– 2024.– Vol. **12**.– Pp. 166843–166854. ↑45

[24] Улизько М. В., Ватьян А. С., Гусарова Н. Ф., Добренко Н. В. *Применение больших языковых моделей для юридических экспертиз* // Экономика. Право. Инновации.– 2025.– № 1.– С. 57–68. ↑46

[25] Parmar M., Patel N., Varshney N., Nakamura M., Luo M., Mashetty S., Mitra A., Baral C. *Towards systematic evaluation of logical reasoning ability of large language models.*– 2024.– 29 pp. arXiv 2404.15522 ↑46

[26] Jiang C., Yang X. *Agentsbench: A multi-agent LLM simulation framework for legal judgment prediction* // Systems.– 2025.– Vol. **13**.– No. 8.– id. 641.– 21 pp. URL ↑46

[27] Жучкова С. В., Девятников В. Ю., Казун А. П., Белов М. Д., Сидорова О. И. *Тексты судебных приговоров как источник данных для эмпирических исследований права в России* // Мониторинг общественного мнения: экономические и социальные перемены.– 2025.– № 2.– С. 170–192. ↑47

Поступила в редакцию 24.12.2025;
одобрена после рецензирования 03.02.2026;
принята к публикации 17.02.2026;
опубликована онлайн 23.02.2026.

Рекомендовал к публикации *к.т.н. Е. П. Куршев*

Информация об авторах:

Юрий Петрович Сердюк

старший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна, научные интересы: параллельное программирование, формальные исчисления процессов, системы типов.

0000-0003-2916-2102
e-mail: *Yuri@serdyuk.botik.ru*

Наталья Александровна Власова

младший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна, научные интересы: компьютерная лингвистика, автоматическая обработка естественного языка, корпусная лингвистика.

0000-0002-7843-6870

e-mail: *nathalie.vlassova@gmail.com*

Седа Рубеновна Момот

младший научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна, научные интересы: компьютерная лингвистика, автоматическая обработка естественного языка.

0000-0002-6097-6545

e-mail: *morlot@mail.ru*

Елена Анатольевна Сулейманова

научный сотрудник Исследовательского центра искусственного интеллекта ИПС им. А. К. Айламазяна, научные интересы: компьютерная лингвистика, автоматическая обработка естественного языка.

0000-0002-0792-9651

e-mail: *yes@helen.botik.ru*

Авторы внесли равный вклад в подготовку публикации.

Декларация об отсутствии личной заинтересованности: *благополучие авторов не зависит от результатов исследования.*

EN RU PROGRAM SYSTEMS: THEORY AND APPLICATIONS ISSN 2079-3316
Research Article ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

UDC 004.89:343.15
10.25209/2079-3316-2026-17-1-21-56

Legal judgement analysis using large language models

Yuri **Serdyuk**[1], Natalia **Vlasova**[2], Seda **Momot**[3], Elena **Suleymanova**[4]
[1–4] Ailamazyan Program Systems Institute of RAS, Ves'kovo, Russia

Abstract. The article examines the use of the latest-generation large language models (LLMs) — such as ChatGPT, Grok, DeepSeek, GigaChat, and YandexGPT — for analyzing legal judgments. The analysis involved civil, administrative, and criminal cases. A dataset of legal judgments was compiled from the database of judicial and regulatory acts of the Russian Federation, the official portal of the Moscow courts of general jurisdiction, and the website of the Russian Agency for Legal and Judicial Information. Several types of large model tests were proposed and implemented, ground-truth selection principles were outlined, and queries (prompts) were formulated. The models were tested on their ability to predict appellate decisions, map crime descriptions to law articles, and evaluate decisions of multiple judicial authorities in a single case. The ability of the models to make their own consistent decisions was also examined. Testing showed that the correct prediction rate of LLMs on real-world juducial decisions rarely surpasses 50%. A brief overview of recent publications on the use of AI in legal practice is provided. (*In Russian*).

Key words and phrases: large language models, LLM, legal judgements, dataset, prompt, AI in law, LegalAI

2020 *Mathematics Subject Classification*: 68T37; 91F99, 68T05

For citation: Yuri Serdyuk, Natalia Vlasova, Seda Momot, Elena Suleymanova. *Legal judgement analysis using large language models*. Program Systems: Theory and Applications, 2026, **17**:1(70), pp. 21–56. (*In Russ.*).
https://psta.psiras.ru/read/psta2026_1_21-56.pdf

References

[1] X. Cshajge. "Automatization of legal departments: survey results LegalInsight", *LegalInsight*, 2023, no. 10 (126), pp. 18–25 (in Russian).

[2] *The experience of law professors who already use AI: real cases and tasks*, 2025 (in Russian).

[3] R. J. Couture. *The impact of artificial intelligence on law firms' business models*, February 25, Center on the Legal Profession, Harvard Law School, 2025.

[4] S. White. *How AI is reshaping the future of legal practice*, The Law Society, 2024.

[5] *Legal innovation: Seizing the future or falling behind?* Wolters Kluwer 2024 Future Ready Lawyer Survey Report, 2024, 18 pp.

[6] *Gartner identifies the Top 6 use cases for generative AI in legal departments*, February 19, 2025.

[7] A. Alibekov, A. Matenkov, V. Bolshakov, G. Mukhtarova, A. Migal, A. Muryshev, A. Kozachenko, N. Mikhaylovskiy. "RuTaR — A dataset in Russian for reasoning about taxes", *Proceedings of the International Conference "Dialogue 2025"* (Moscow, Russia, 23–25 April 2025), Computational Linguistics and Intellectual Technologies, vol. **23**, 2025, pp. 12–27.

[8] O. N. Gorbunov. "From patterns to precedents: application of large language models in legal practice", *Sud'ya*, 2026, no. 1 (in Russian).

[9] Y. Fan, J. Ni, J. Merane, Y. Tian, Y. Hermstrüwer, Y. Huang, M. Akhtar, E. Salimbeni, F. Geering, O. Dreyer, D. Brunner, M. Leippold, M. Sachan, A. Stremitzer, Ch. Engel, E. Ash, J. Niklaus. *LEXam: benchmarking legal reasoning on 340 law exams*, 2025, 38 pp. arXiv 2505.12864

[10] J. Shen, J. Xu, H. Hu, L. Lin, F. Zheng, G. Ma, F. Meng, J. Zhou, W. Han. "A law reasoning benchmark for LLM with tree-organized structures including factum probandum, evidence and experiences", *Findings of the Association for Computational Linguistics: ACL 2025*, The 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025) (Vienna, Austria, July 27 — August 1 2025), ACL, 2025, ISBN 9798331323936, pp. 17252–17274. arXiv 2503.00841

[11] M. Medvedeva, P. McBride. "Legal judgment prediction: if you are going to do it, do it right", *Proceedings of the Natural Legal Language Processing Workshop* (Singapore, 7 December 2023), ACL, 2023, ISBN 9781713886044, pp. 73–84.

[12] M. Sesodia, A. Petrova, J. Armour, T. Lukasiewicz, O.-M. Camburu, P. Dokania, P. Torr, C. Schröder de Witt. *AnnoCaseLaw: a richly-annotated dataset for benchmarking explainable legal judgment prediction*, 2025, 15 pp. arXiv 2503.00128

[13] S. K. Nigam, D. P. Balaramamahanthi, S. Mishra, N. Shallum, K. Ghosh, A. Bhattacharya. "NyayaAnumana and INLegalLlama: the largest Indian legal judgment prediction dataset and specialized language model for enhanced decision analysis", *Proceedings of the 31st International Conference on Computational Linguistics* (Abu Dhabi, United Arab Emirates, 19–24 January 2025), ACL, 2025, ISBN 979-8-3313-1356-2, pp. 11135–11160.

[14] M. B. Kmainasi, A. E. Shahroor, A. Al-Ghraibah. *Can large language models predict the outcome of judicial decisions?* 2025, 6 pp. arXiv 2501.09768

[15] E. A. Posner, S. Saran. *Judge AI: assessing large language models in judicial decision-making*, Coase-Sandor Institute for Law & Economics Research Paper Series, No 25-03, Coase-Sandor Institute for Law & Economics, 2025, 41 pp.

[16] A. P. Kazun. "Can artificial intelligence predict judicial decisions? A systematic review of international research", *Monitoring obshhestvennogo mneniya: ekonomicheskie i social'nye peremeny*, 2024, no. 5, pp. 100–122 (in Russian).

[17] M. Mazukov. *Neural networks assessed Dolina's case*, 2025 (in Russian).

[18] M. Mazukov. *Artificial intelligence for lawyers: How AI assisted in overturning a lower court decision*, 2025 (in Russian).

[19] S. K. Nigam, B. D. Patnaik, S. Mishra, A. V. Thomas, N. Shallum, K. Ghosh, A. Bhattacharya. "NyayaRAG: realistic Legal Judgment Prediction with RAG under the Indian common law system", 2025, 18 pp. arXiv 2508.00709

[20] P. Laban, H. Hayashi, Y. Zhou, J. Neville. *LLMs get lost in multi-turn conversation*, 2025, 36 pp. arXiv 2505.06120

[21] H. Min, B. Noh. "TRACS-LLM: LLM-based traffic accident criminal sentencing prediction focusing on imprisonment, probation, and fines", *Artificial Intelligence and Law*, 2025, 22 pp.

[22] B. Wei, Y. Yu, L. Gan, F. Wu. "An LLMs-based neuro-symbolic legal judgment prediction framework for civil cases", *Artificial Intelligence and Law*, 2025, 35 pp.

[23] X. Wang, X. Zhang, V. Hoo, Z. Shao, X. Zhang. "LegalReasoner: A multi-stage framework for legal judgment prediction via large language models and knowledge integration", *IEEE Access*, **12** (2024), pp. 166843–166854.

[24] M. V. Uliz'ko, A. S. Vat'yan, N. F. Gusarova, N. V. Dobrenko. "Application of large language models for legal expertise", *Ekonomika. Pravo. Innovacii*, 2025, no. 1, pp. 57–68 (in Russian).

[25] M. Parmar, N. Patel, N. Varshney, M. Nakamura, M. Luo, S. Mashetty, A. Mitra, C. Baral. "Towards systematic evaluation of logical reasoning ability of large language models", 2024, 29 pp. arXiv 2404.15522

[26] C. Jiang, X. Yang. "Agentsbench: A multi-agent LLM simulation framework for legal judgment prediction", *Systems*, **13**:8 (2025), id. 641, 21 pp.

[27] S. V. Zhuchkova, V. Yu. Devyatnikov, A. P. Kazun, M. D. Belov, O. I. Sidorova. "Texts of court verdicts as a data source for empirical legal studies in Russia", *Monitoring obshhestvennogo mneniya: ekonomicheskie i social'nye peremeny*, 2025, no. 2, pp. 170–192 (in Russian).