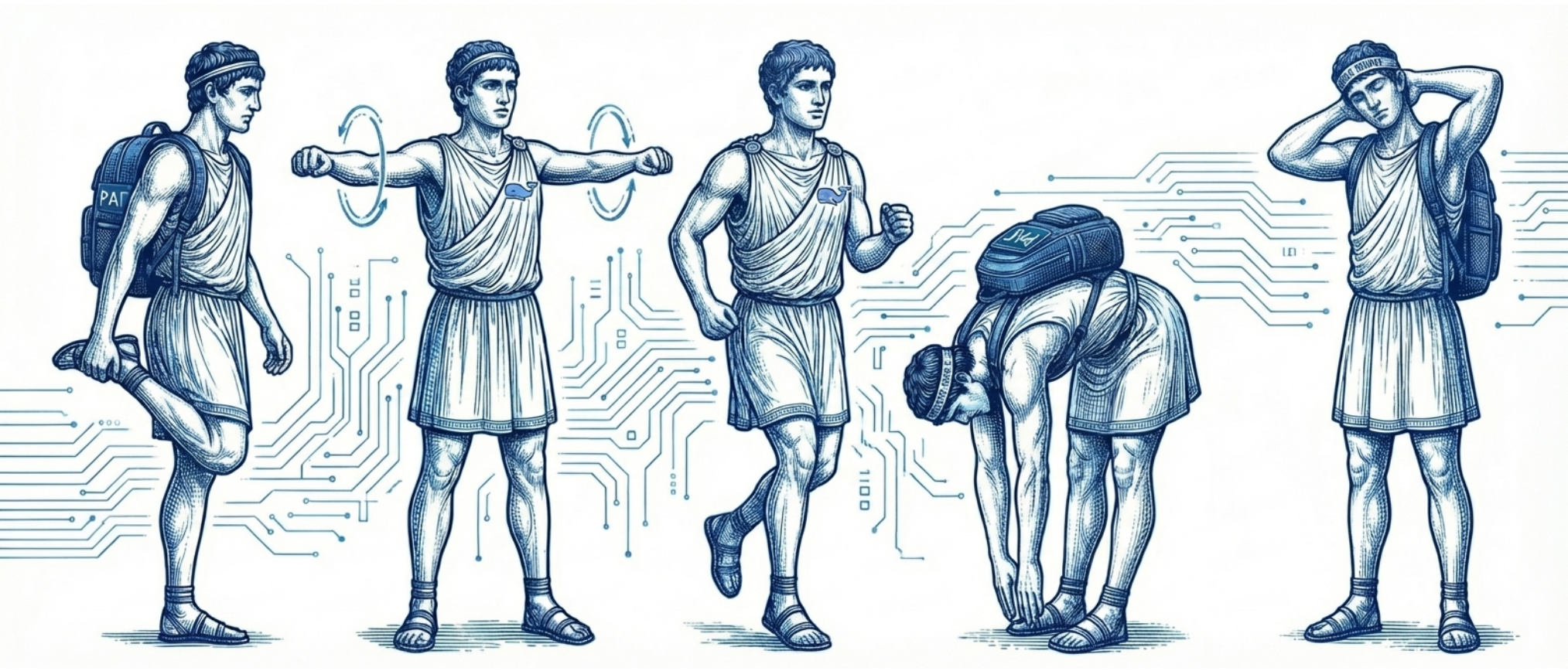


Делай Bench

эксперимент по проведению слепого human-eval бенчмарка больших языковых моделей для юридических задач

Екатерина Якуненко

Автор Telegram-канала «Делай RAG»

декабрь 2025

Предисловие

Решение провести этот бенчмарк (скорее даже эксперимент по его проведению) родилось спонтанно: возможность запустить на сильной видеокарте open-source модель, представленную сообществу Нейросети | ilovedocs Кенаном Коюшовым, быстро переросла в затею сравнить работу этой модели с другими сервисами. За пару дней удалось собрать вопросы и найти оценщиков, ещё за пару — подготовить для оценщиков данные, подождать неделю ответов и ещё за три дня посчитать результаты и собрать их в этот отчёт. Очень хотелось успеть до Нового года, чтобы разработчикам было о чём подумать в праздники :)

В каком-то смысле всё сделано «на коленке», поэтому методология, конечно, не совершенна. Но даже в условиях несовершенной методологии удалось получить, как мне кажется, очень интересные результаты.

Хотелось бы, чтобы этот опыт был не последним, а юридическое сообщество узнало о таком формате оценки LLM, увидело в нём ценность, и у нас, как у пользователей, начала формироваться культура правильных ожиданий от способностей представленных на рынке нейросетевых продуктов. А у разработчиков таких продуктов было понимание того, что пользователи ожидают убедительных и объективных оценок, а не только маркетинговых заявлений.

Об организаторе бенчмарка

Меня зовут Екатерина Якуненко, я более 6 лет работаю юристом, и с июня 2025 активно погружаюсь в разные аспекты использования больших языковых моделей в работе юристов.

Мои основные проекты:

- Telegram-канал «Делай RAG» и различные публикации на тему RAG-технологии в юридическом домене
- Telegram-бот «А юристы смотрели?», умеющий проверять рекламные креативы на соответствие ФЗ «О рекламе», используя базу знаний из решений ФАС по RAG-технологии, и открытый датасет с размеченной практикой ФАС по рекламным делам, который можно скачать здесь или здесь
- материал «Как испытать от вайб-кодинга радость» о моей практике использования нейросетей для разработки без знания языков программирования

Содержание
(строки кликабельны)

РАЗДЕЛ 1

Резюме эксперимента

Главная цель: Провести независимое слепое сравнительное тестирование специализированных LLM-сервисов для юристов и общедоступных больших языковых моделей общего назначения на задачах, приближённых к реальной юридической практике.

Задачи исследования:

- ✦ Сравнить эффективность моделей общего назначения (*DeepSeek*) и специализированных юридических решений (*Нейроюрист*, *АйЮрист*).
- ✦ Проверить гипотезу о ценности RAG-систем (моделей с доступом к базе знаний) над «чистыми» языковыми моделями.
- ✦ Сравнить результативность машинного обучения и использования RAG-технологии.
- ✦ Проверить валидность методологии экспертной оценки (*human evaluation*), учитывающую уровень уверенности юристов-оценщиков в своих ответах.

Неожиданные (ли?) открытия

1. Ожидаемо, ни одна модель не доминирует во всех метриках. Но наиболее надёжным в среднем оценщики сочли «думающий» DeepSeek.

2. Модели с доступом к актуальной базе знаний (т.е. с RAG) показывают радикально лучшие результаты в **профильных** темах.

3. «Красноречие» ≠ качество. Модель, которая чаще всего получает первые места (DeepSeek-V3), проигрывает при проверке экспертами с высокой уверенностью. Убедительность текста маскирует юридические неточности.

4. Модель, занявшая последнее место в общем зачёте (Kenan+RAG), в тех случаях, когда она давала ответ, способна показать результаты на уровне лидеров рынка.

РАЗДЕЛ 2

Дизайн эксперимента

2.1. Тестируемые LLM

Большие языковые модели общего назначения:

- *DeepSeek-R1* (reasoner) — модель с режимом «глубокого размышления», оптимизированная для сложных логических задач. Браузерная версия https://chat.deepseek.com/, доступная по кнопке «Deep Think» в интерфейсе. Использовалась без включённого поиска в Интернете.
- *DeepSeek-V3* — универсальная модель последнего поколения с акцентом на качество генерации текста. Браузерная версия https://chat.deepseek.com/. Использовалась без включённого поиска в Интернете.

Выбор сделан в пользу именно DeepSeek, так как он доступен для пользователей из России без средств обхода блокировок.

Специализированные юридические решения:

- *Нейроюрист* — российский сервис от Яндекс с интегрированной базой законодательства от СПС «Гарант» по RAG-технологии по ряду отраслей права. Браузерная версия https://neurolegal.ya.ru/.
- *Kenan (Base)* — open-source модель Ken1.0-67B (дообученный на российской нормативной и правоприменительной базе Qwen 2). Запущена на двух видеокартах NVIDIA H100.
- *Kenan+RAG* (сервис АйЮрист) — та же модель с подключённой системой поиска по самостоятельно собранной нормативной и правоприменительной базе знаний по RAG-технологии. Браузерная версия https://chat.айюрист.рф.

2.2. Принцип слепого тестирования

В основе бенчмарка лежит методология *blind testing* — «слепого» тестирования практикующими юристами. **Ключевые элементы дизайна:**

- *Анонимизация ответов* — все ответы моделей были обезличены — эксперты не знали, какой текст сгенерирован какой системой.
- *Рандомизация порядка* — последовательность ответов в каждом листе Excel-формы опроса была случайной, чтобы исключить эффект позиции («первый ответ всегда кажется лучше») или догадку эксперта об отвечающей LLM, что могло бы привести к предвзятости оценки.
- *Ранжирование, а не абсолютные оценки* — вместо субъективных баллов («насколько хорош ответ по шкале от 1 до 10») эксперты распределяли ответы по местам **от 1-го до 5-го**. Это заставляет делать выбор и минимизирует инфляцию оценок.

2.3. Состав оценщиков

В оценке приняли участие **11 юристов** различной специализации. Количество экспертов в целом невелико с точки зрения классической статистики, однако каждый эксперт оценивал от 1 до 5 вопросов в зависимости от своей компетенции, и суммарно было получено достаточное количество (**155**) независимых оценок для статистического анализа с поправкой на уровень уверенности эксперта в своём ответе.

2.4. Критерии оценки

Экспертам была поставлена задача ранжировать ответы, учитывая комплекс факторов:

- *Логичность и структура изложения* — насколько понятно и последовательно построен ответ.
- *Юридическая корректность* — правильность ссылок на НПА, соответствие актуальной практике.
- *Практическая полезность* — помогает ли ответ решить поставленную в вопросе задачу.
- *Отсутствие «галлюцинаций»* — нет ли в ответе выдуманных норм, несуществующих статей или ложных утверждений.

2.5. Уровень уверенности в ответе

Чтобы минимизировать искажения результатов от «одноуровневости» оценок экспертов с разным бэкграундом и специализацией, был введён параметр **самооценки уверенности** оценщика в ответе.

После каждого вопроса эксперт выставлял себе балл от 1 до 5, где:

- **5 баллов** = «Я глубоко разбираюсь в этой теме, точно знаю правильный подход к решению задачи»
- **1 балл** = «Оцениваю только по логике и здравому смыслу, это не моя область»

Этот параметр позволил в дальнейшем применить **взвешивание** результатов: голос эксперта с высокой уверенностью влияет на итоговый рейтинг сильнее, чем голос «дженералиста».

2.6. Набор тестовых вопросов

Для тестирования оценщиками и организатором бенчмарка были составлены **10 вопросов**, охватывающих различные отрасли права.

Вопросы были сгруппированы по тематическим кластерам для последующего анализа специализации моделей:

- Семейное право: Q1, Q10
- Общие вопросы частного права: Q2, Q3, Q4
- Управление общедомовым имуществом: Q5, Q6
- Коммерческое право: Q7, Q8, Q9

Q1	Ты опытный юрист, специализирующийся на семейном праве Российской Федерации. Задача 1. Ответь на следующие вопросы: 1) Найди нормативно-правовое обоснование расчета задолженности по алиментам и индексации. Может ли быть применена индексация после увеличения размера алиментов апелляцией? Задача 2. Рассчитай задолженность по алиментам с учетом контекста. Контекст: алименты с 12.05.2025 выплачиваются в размере 1 минимального прожиточного минимума на ребенка, с 06.12.2025 размер алиментов увеличен до 3 минимальных прожиточных минимумов на ребенка с 12.05.2025. Второй суммой рассчитай сумму индексации.
Q2	Ты опытный юрист, специализирующийся в вещном праве Российской Федерации. Представь анализ правового обоснования и условий регистрации обратного перехода права собственности на недвижимое имущество при расторжении договора купли-продажи недвижимости по законодательству Российской Федерации. Выяви ключевое отличие в ситуациях, когда договор исполнен в полном объеме надлежащим образом или ненадлежащим образом.
Q3	Ты опытный юрист, специализирующийся в частном праве Российской Федерации. Дай аргументированный ответ на вопрос, можно ли понудить к исполнению в натуре обязанности по внесению / восполнению обеспечительного платежа?
Q4	Ты опытный юрист, специализирующийся в обязательственном и коммерческом праве Российской Федерации. Дай аргументированный ответ на вопрос, вправе ли подрядчик по договору строительного подряда предъявить исковое заявление к заказчику о понуждении к приемке выполненных работ?
Q5	Ты опытный юрист, специализирующийся в управлении многоквартирными жилыми домами по законодательству Российской Федерации. Дай исчерпывающую инструкцию для следующей задачи: что необходимо сделать для установки кондиционера на фасад многоквартирного дома?
Q6	Ты опытный юрист, специализирующийся в управлении многоквартирными жилыми домами по законодательству Российской Федерации. Дай исчерпывающую инструкцию для следующей задачи: что необходимо сделать для демонтажа подоконного пространства в многоквартирном доме?

Q7	Ты опытный юрист, специализирующийся на законодательстве о защите прав потребителей Российской Федерации. Предоставь продавцу уникальных спортивных (match-worn профессиональными спортсменами) товаров, организующему он-лайн аукционы, консультацию по следующему вопросу: имеет ли победитель аукциона отказаться от выигранного им товара? Какие нормы российского законодательства регулируют этот вопрос? Какими способами продавец может защитить свои интересы и ограничить потребителей в злоупотреблении возможностью отказываться от товаров и срывать аукционы, кроме задатка?
Q8	Ты опытный юрист, специализирующийся на вопросах добросовестной конкуренции по законодательству Российской Федерации. Предоставь продавцу детских переносок-слингов консультацию о том, каким образом ей необходимо вести свои социальные сети и делать обзоры нефизиологичных переносок, чтобы в полной мере описывать их недостатки и риски использования, снизив до максимума для себя риск претензий от конкурентов о защите деловой репутации и недобросовестной конкуренции? Какими нормативными правовыми актами, стандартами, регламентами регулируется вопрос безопасности и физиологичности детских переносок?
Q9	Ты опытный юрист, специализирующийся на лицензировании фармацевтической деятельности по законодательству Российской Федерации. Дай аргументированный ответ на вопрос, имеет ли право фармацевтическая компания взять в аренду склад для транзита лекарственных препаратов внутри России, не внося при этом изменений в имеющуюся лицензию? Опиши возможные риски, связанные с невнесением изменений, если они необходимы, а также предложи возможные варианты действий для компании, если внесение изменений в лицензию нецелесообразно.
Q10	Супруги заключили брачный договор, когда у них уже были обязательства перед кредиторами, кредиторов не уведомили о заключении брачного договора. Через 2 года в отношении супруга началась процедура банкротства. В реестр кредиторов должника включены кредиторы (далее – старые кредиторы), обязательства перед которыми возникли до заключения брачного договора. 18.06.2025 Верховный Суд РФ выпустил Обзор по банкротству граждан, где указал в пункте 41, что брачный договор, для старых кредиторов не противопоставляется. Вопросы: 1) Есть ли необходимость управляющему оспаривать брачный договор, если в реестре только старые кредиторы? 2) Как быть с оспариванием сделок между должником и супругой? Между ними продолжает действовать раздельный режим? Например, должник вносил на счет ИП супруги крупные суммы, можно ли это оспорить как безвозмездную сделку третьему лицу?

2.7. Ограничения эксперимента

Для корректной интерпретации результатов необходимо учитывать следующие ограничения исследования:

- *Формат «один вопрос — один ответ».* Тестировался статический сценарий, где модель получает запрос и должна сразу дать финальный ответ.

  В реальной практике юрист может в ходе диалога с моделью уточнять детали, загружать документы, задавать уточняющие вопросы и отвечать на задаваемые LLM вопросы. Модель *Kenan+RAG* оказалась настроенной именно на диалоговый режим, что могло существенно повлиять на её результаты.
- *Выборка экспертов*. 11 юристов — это достаточно для выявления тенденций, но недостаточно для статистически «железобетонных» выводов. Один строгий или, наоборот, лояльный оценщик может заметно повлиять на картину.
- *Неравномерная сложность вопросов.* Часть вопросов оказалась «понятной» для большинства экспертов (классическое гражданское право), часть — узкоспециальной (фармацевтика, технические регламенты). Это отразилось на разбросе оценок.
- *Воспроизводимость результатов.* Тестирование проводилось в конкретный момент времени (середина декабря 2025). Модели постоянно обновляются, законодательство меняется — результаты могут не воспроизводиться через полгода.

РАЗДЕЛ 3

Ход эксперимента и аналитика

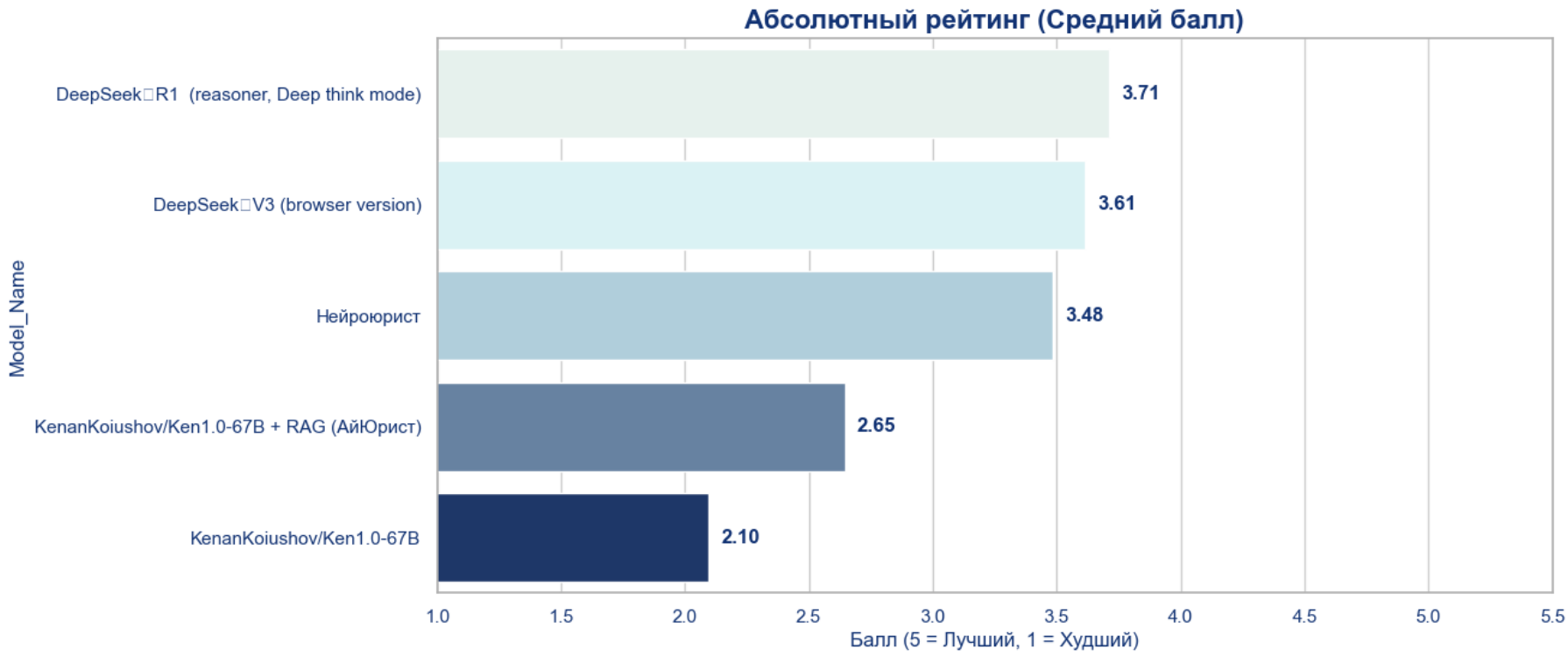
3.1. Этап 1: Базовые метрики

После сбора всех оценок данные были приведены к единому виду. Поскольку в анкетах 1-е место означало лучший ответ, а 5-е — худший, шкала была инвертирована: теперь 5 баллов = лучший результат, 1 балл = худший. Это проще для подсчётов и интуитивно понятнее.

Метрика 1: Средний балл (Mean Score)

Простое среднее арифметическое всех оценок, полученных моделью. Показывает «среднюю температуру по больнице» — общий уровень модели без учёта нюансов.

Ограничения: не учитывает, что один оценщик может быть строгим (ставит всем 4-5), а другой — лояльным (ставит 1-2). Также не учитывает экспертизу оценщика в конкретной теме.



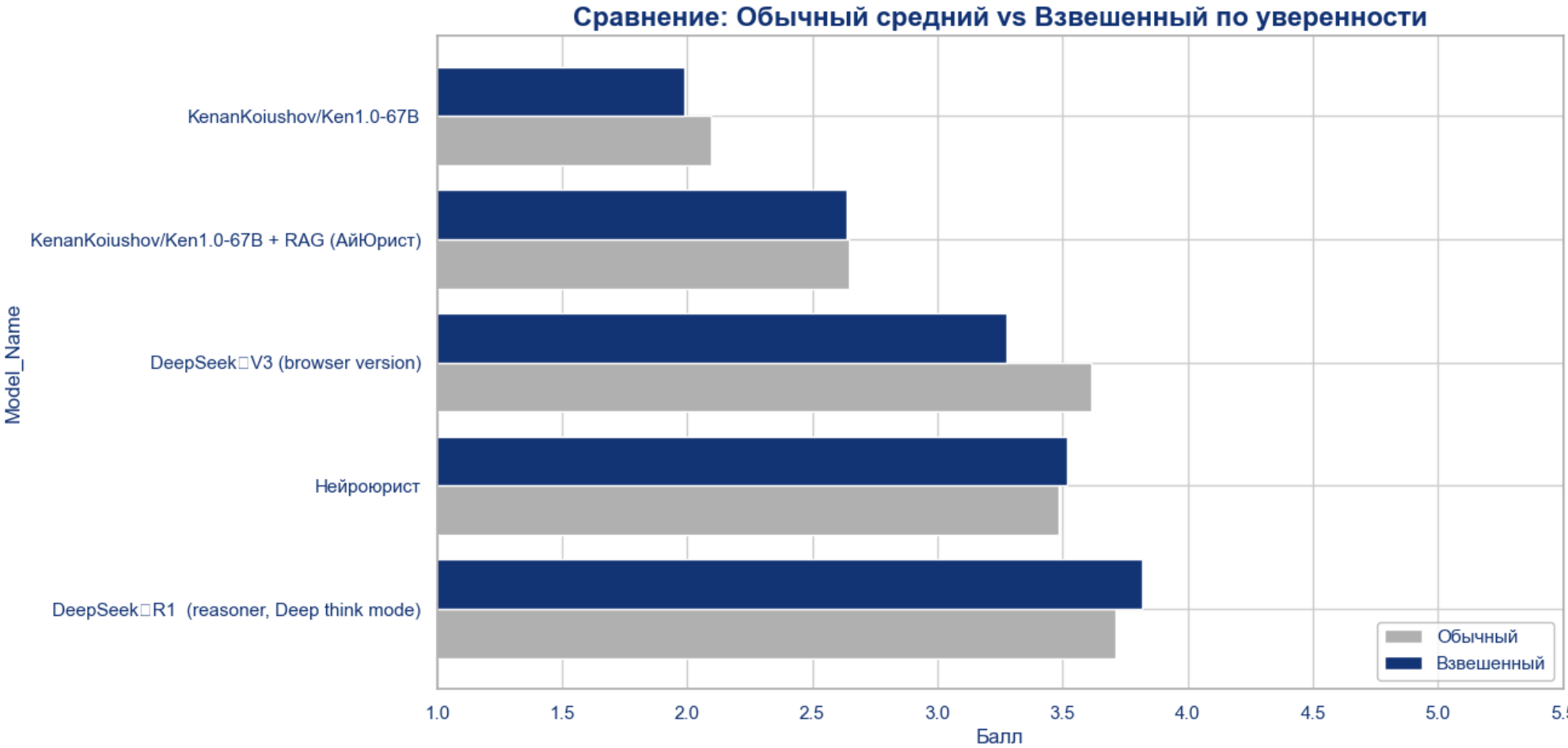
Чётко выделяются две группы: «высшая лига» (*DeepSeek* и *Нейроюрист* с баллами выше 3.4) и «отстающие» (семейство *Kenan* ниже 2.7).

Примечательно, что добавление RAG к базовой модели *Kenan* даёт прирост в 0.55 балла — существенное улучшение, хотя и недостаточное для конкуренции с лидерами.

Метрика 2: Взвешенный рейтинг (Weighted Score)

Средний балл, где каждая оценка умножена на коэффициент уверенности эксперта.

Позволяет «утяжелить» голоса профильных специалистов. Взвешенный рейтинг фильтрует «шум» от оценок, поставленных «по логике», и показывает, какую модель выбирают настоящие эксперты, когда они точно знают правильный ответ.



При переходе от простого среднего к взвешенному происходят значимые сдвиги:

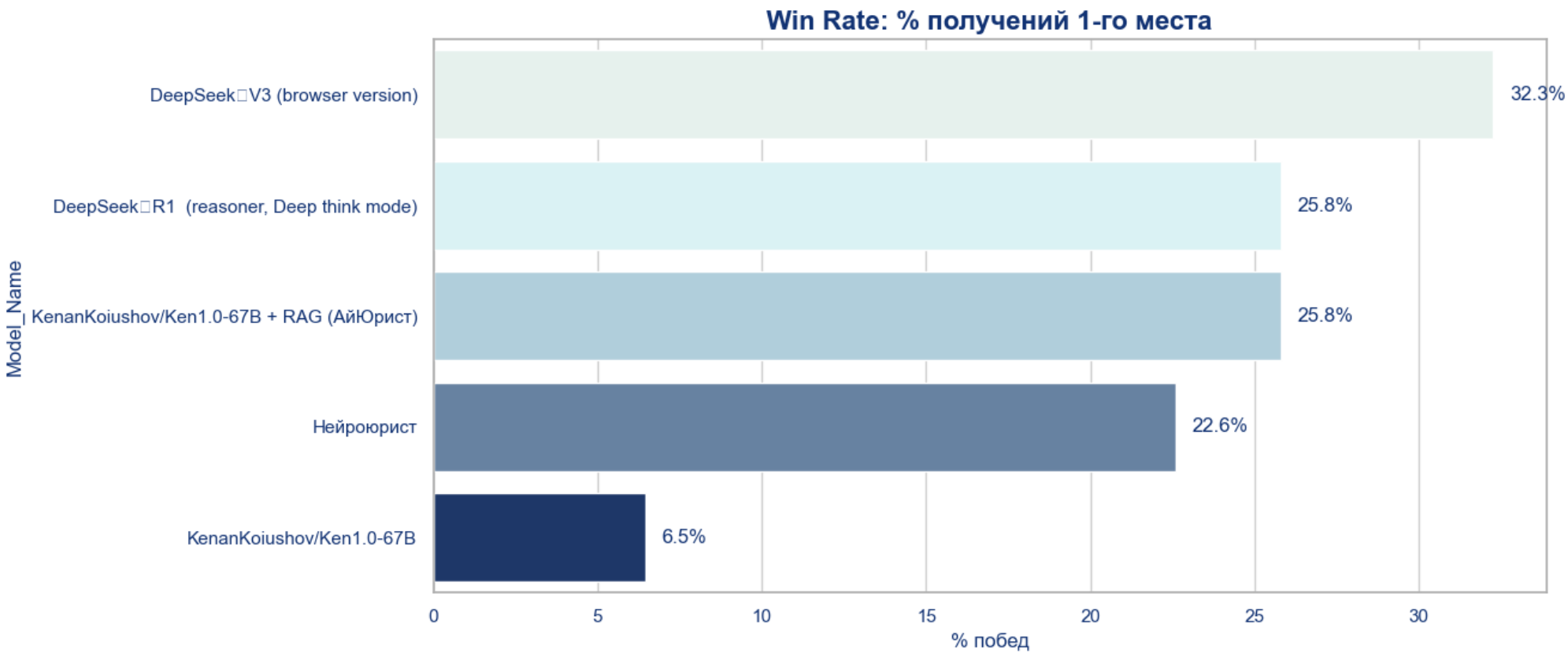
- *DeepSeek-V3* падает с 3.61 до 3.27. Это означает, что высокие баллы модель получала преимущественно от неуверенных оценщиков. Когда эксперты точно знали правильный ответ, они ставили ей ниже.
- *DeepSeek-R1* и *Нейроюрист* растут. Профессионалы с высокой уверенностью предпочитают именно эти модели.

Этот паттерн — важный сигнал. *DeepSeek-V3* пишет убедительно и красиво, но допускает ошибки, которые замечают только специалисты. *DeepSeek-R1*, напротив, может казаться менее «гладкой», но даёт более точные ответы.

Метрика 3: Win Rate (процент первых мест)

Доля случаев, когда модель заняла чистое 1-е место.

Показывает способность модели выдавать «вау-результат» — ответ, который эксперт однозначно признаёт лучшим.



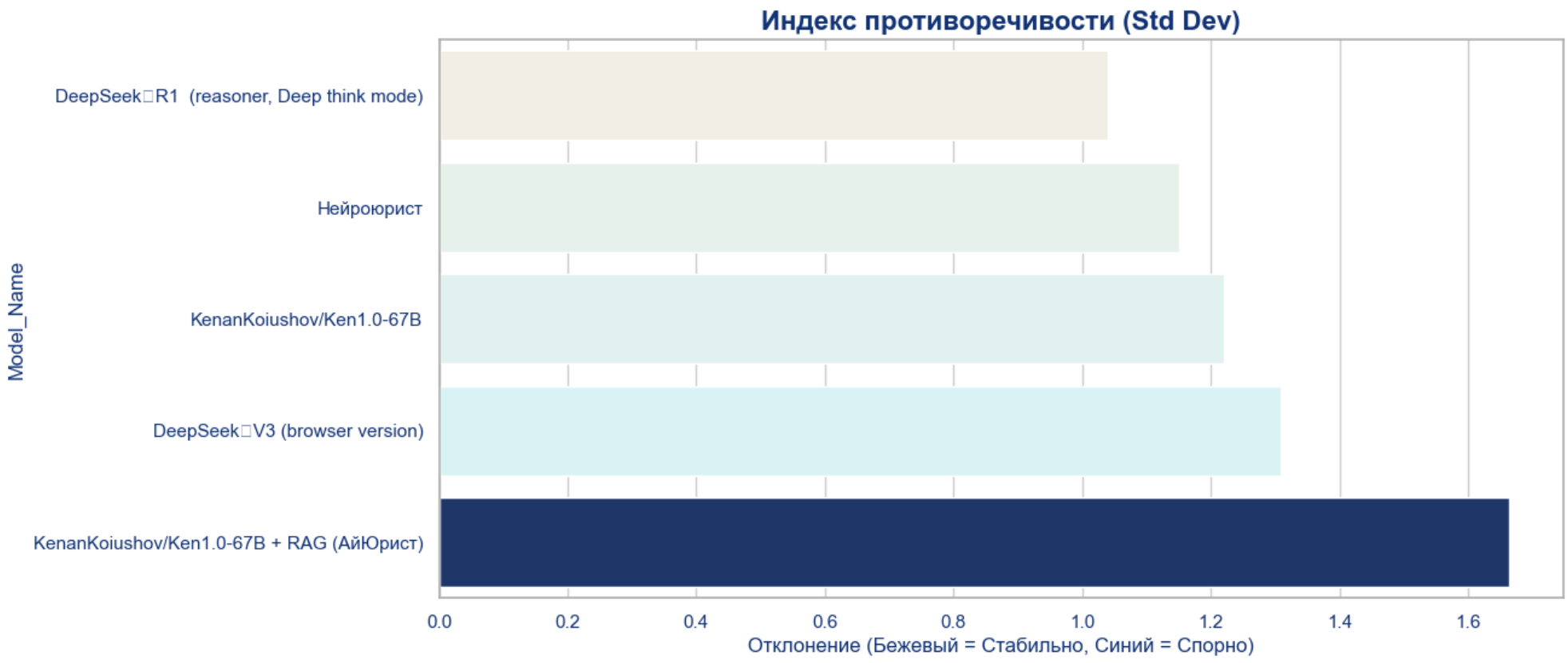
Парадокс DeepSeek-V3: Модель лидирует по количеству первых мест, но проигрывает во взвешенном рейтинге. Модель часто выдаёт идеальный ответ, но ей не хватает стабильности.

Феномен Kenan+RAG: При общем рейтинге на последнем месте модель имеет Win Rate на уровне лидера (25.8%). Это значит, что когда база знаний срабатывает, модель попадает точно в цель. Но когда не срабатывает либо модель отказывается давать ответ без уточняющих вопросов — падает на дно.

Метрика 4: Индекс спорности (Standard Deviation)

Стандартное отклонение — показывает, насколько сильно расходятся оценки одной модели. Позволяет понять, является ли модель «предсказуемой» или «рулеткой».

Низкое отклонение = эксперты единодушны, модель стабильна. Высокое отклонение = одни восхищаются, другие ставят двойки.



DeepSeek-R1 — самая стабильная модель, минимальный разброс оценок.

Kenan+RAG — максимальный разброс, подтверждает гипотезу «рулетки», зависимости от поведения модели: соглашается ли она дать ответ сразу, срабатывает ли её RAG.

*Каждая из 4 метрик посчитана и для каждого из 10 вопросов. Ознакомиться с результатами по каждому вопросу можно в материалах бенчмарка по **этой ссылке** (см. детали в Приложении 2).*

3.2. Этап 2: Анализ согласованности экспертов

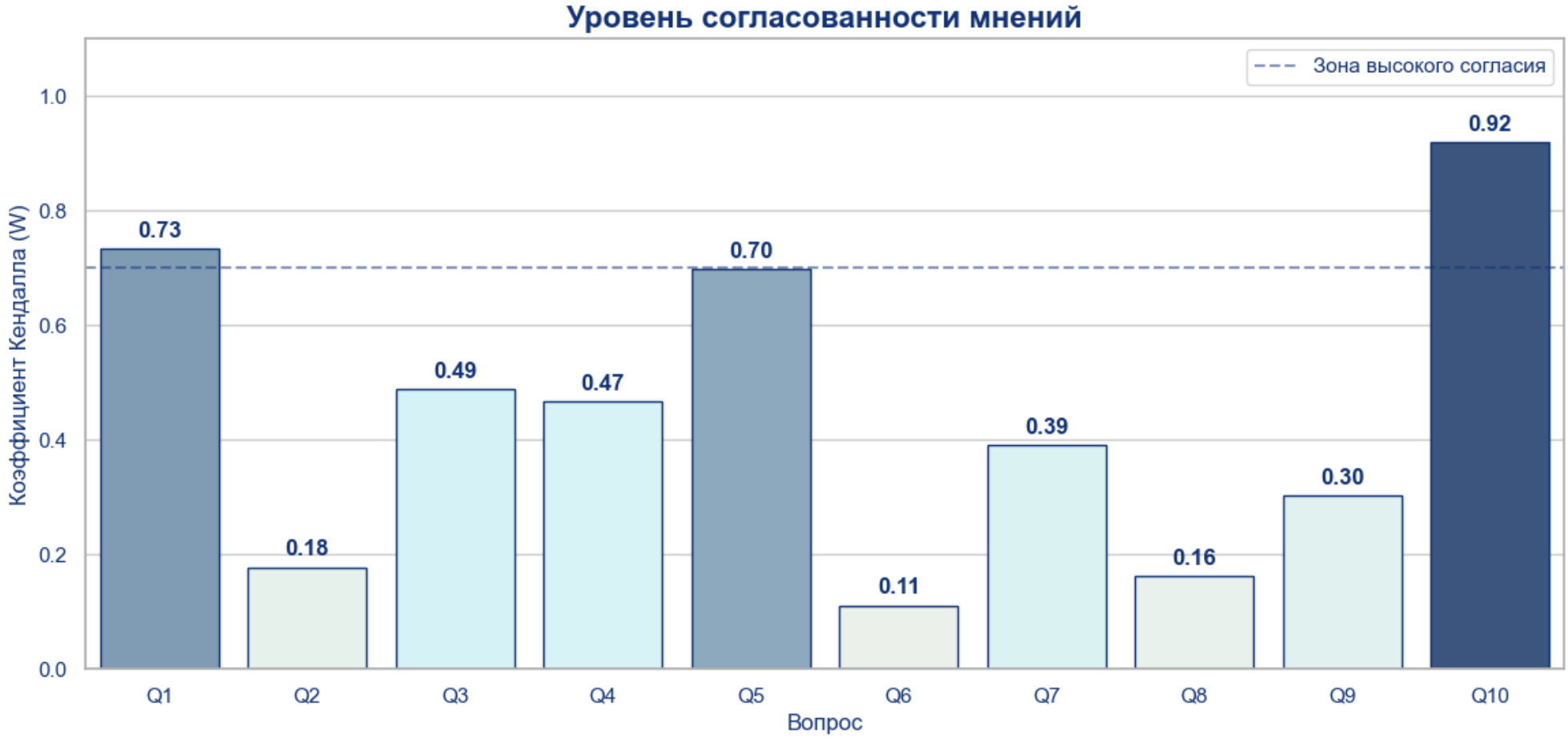
Прежде чем делать выводы о моделях, необходимо понять: насколько объективны сами данные? Если эксперты ставят диаметрально противоположные оценки, это может означать либо спорность вопроса, либо размытость критериев.

Коэффициент согласованности Кендалла (Kendall's W)

Статистический показатель, измеряющий степень единодушия экспертов при ранжировании.

Как интерпретировать:

- ✦ W близко к 1 — эксперты полностью согласны в том, кто лучше, а кто хуже
- ✦ W близко к 0 — полный хаос, каждый ранжирует по-своему
- ✦ W > 0.7 — высокое согласие, результаты надёжны
- ✦ W < 0.4 — низкое согласие, к результатам нужно относиться осторожно



Вопросы чётко разделились на три группы:

«Золотой стандарт» (W > 0.7):

- ✦ Q10 (Банкротство, позиция ВС РФ) — W ≈ 0.9
- ✦ Q1 (Алименты) — W ≈ 0.75
- ✦ Q5 (Кондиционер на фасаде) — W ≈ 0.7

Здесь есть объективная истина: конкретная позиция Верховного Суда, чёткие нормы Семейного кодекса, устоявшаяся практика. Модели ранжировались легко и единодушно.

«Серая зона» (W 0.4–0.6):

✦ Q3 (Обеспечительный платеж) — W ≈ 0.5

✦ Q4 (Понуждение к приёмке) — W ≈ 0.5

Вопросы частного и коммерческого права, где многое зависит от нюансов. Эксперты чувствуют себя уверенно, но сложно однозначно ранжировать ответы разной степени проработки.

«Зона хаоса» (W < 0.3):

✦ Q2 (Обратный переход права собственности) — W ≈ 0.18

✦ Q6 (Демонтаж подоконного блока) — W ≈ 0.1

✦ Q7, Q8, Q9 (узкоспециальные темы) — W 0.2–0.3

По вопросам Q2 и Q6 даже эксперты с высокой уверенностью разошлись во мнениях. Это не проблема моделей — это конфликт правовых позиций.

Корреляция уверенности и согласованности

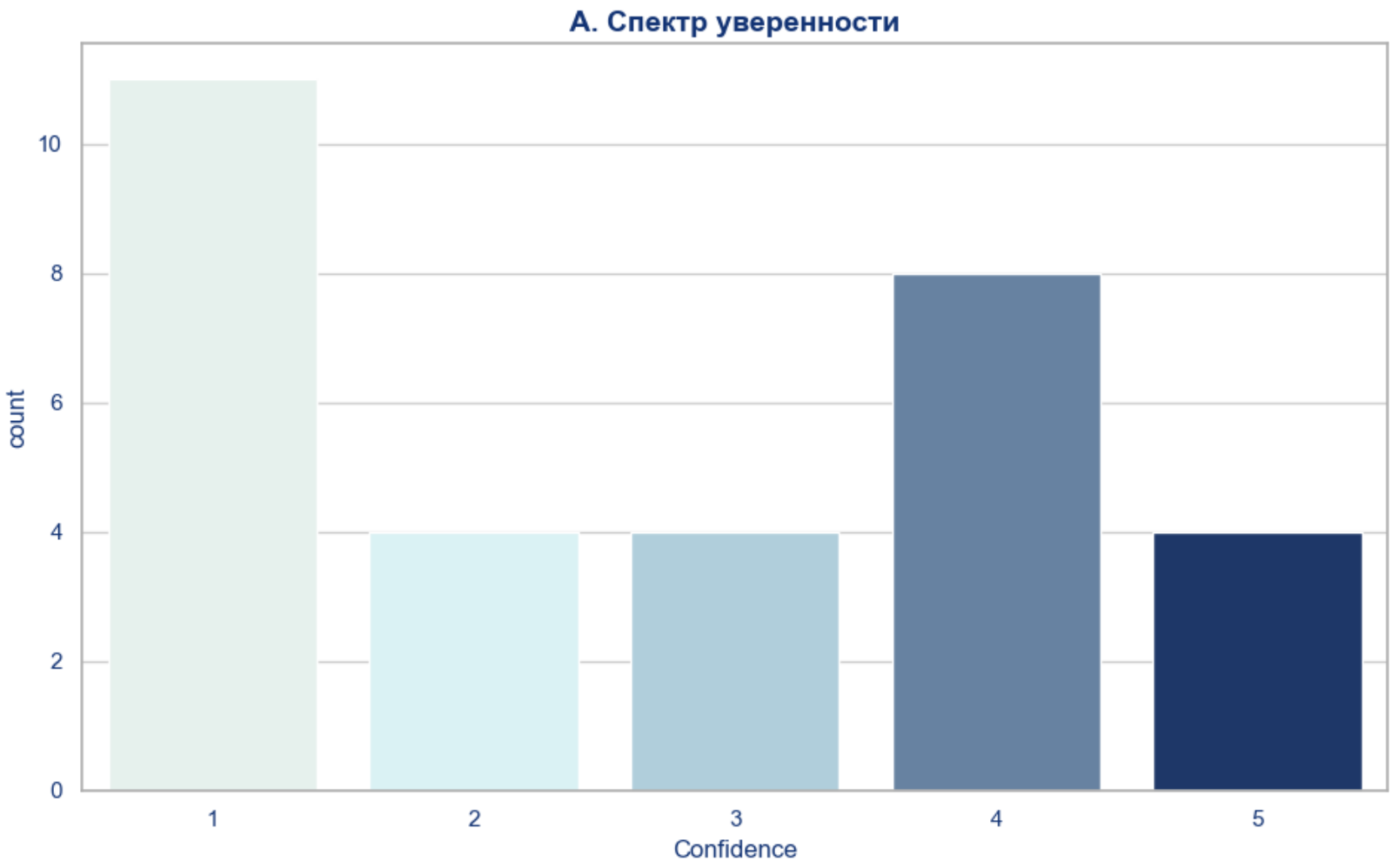
Проверялась гипотеза: влияет ли уверенность экспертов на их единодушие?

Результат: Коэффициент корреляции *r ≈ 0.39* — средняя прямая связь. В целом тенденция подтвердилась: чем увереннее эксперты, тем единодушнее они голосуют. Однако есть яркие аномалии.

Аномалия Q2 и Q6: Высокая уверенность + низкое согласие. Эксперты были уверены в своих знаниях, но ставили диаметрально разные оценки. Это маркер доктринального конфликта — вопросы, по которым в юридическом сообществе нет единой позиции.

Результаты по Q2 и Q6 следует интерпретировать с осторожностью. Оценки там зависели не столько от качества модели, сколько от правовой позиции конкретного эксперта.

3.3. Этап 3: Анализ профилей оценщиков



Общая картина: средняя уверенность по всему проекту составила **2.68 из 5.0** — ниже нейтральной тройки. Это важный показатель: *бенчмарк оказался сложным*, вопросы требовали узкой экспертизы.

Профили экспертов. Оценщики чётко разделились на три группы:

1. *Эксперты-специалисты* (средняя уверенность *3.8–5.0*): оценивали преимущественно «свои» темы, их голоса — самые ценные.

2. *Осторожные критики* (средняя уверенность *~2.7*): подходили к оценке взвешенно, не считали себя экспертами во всём.

3. *Честные дженералисты* (средняя уверенность *1.0–2.25*): открыто признавали: «Это не моя тема». Их оценки по непрофильным вопросам не должны иметь большого веса — и благодаря взвешиванию не имеют.

Самые сложные вопросы (по уровню уверенности экспертов):

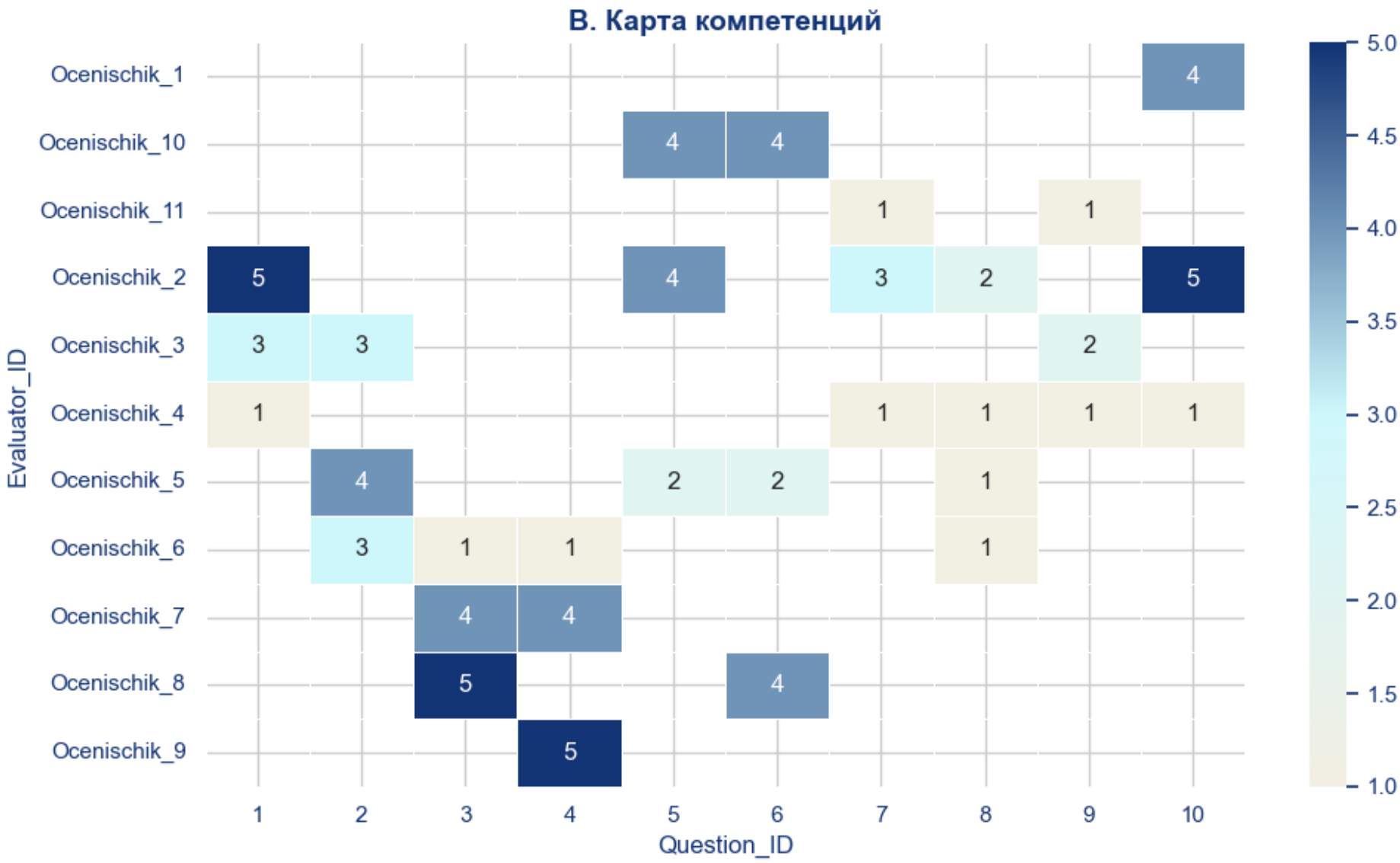
✦ Q8 (Детские переноски, реклама): 1.25

✦ Q9 (Фармацевтика): 1.33

✦ Q7 (Аукционы): 1.67

Самые понятные вопросы:

Q2, Q3, Q4 (классическое гражданское право): 3.33



3.4. Этап 4: Уточняющие эксперименты

В ходе анализа обнаружились аномалии, потребовавшие отдельного изучения.

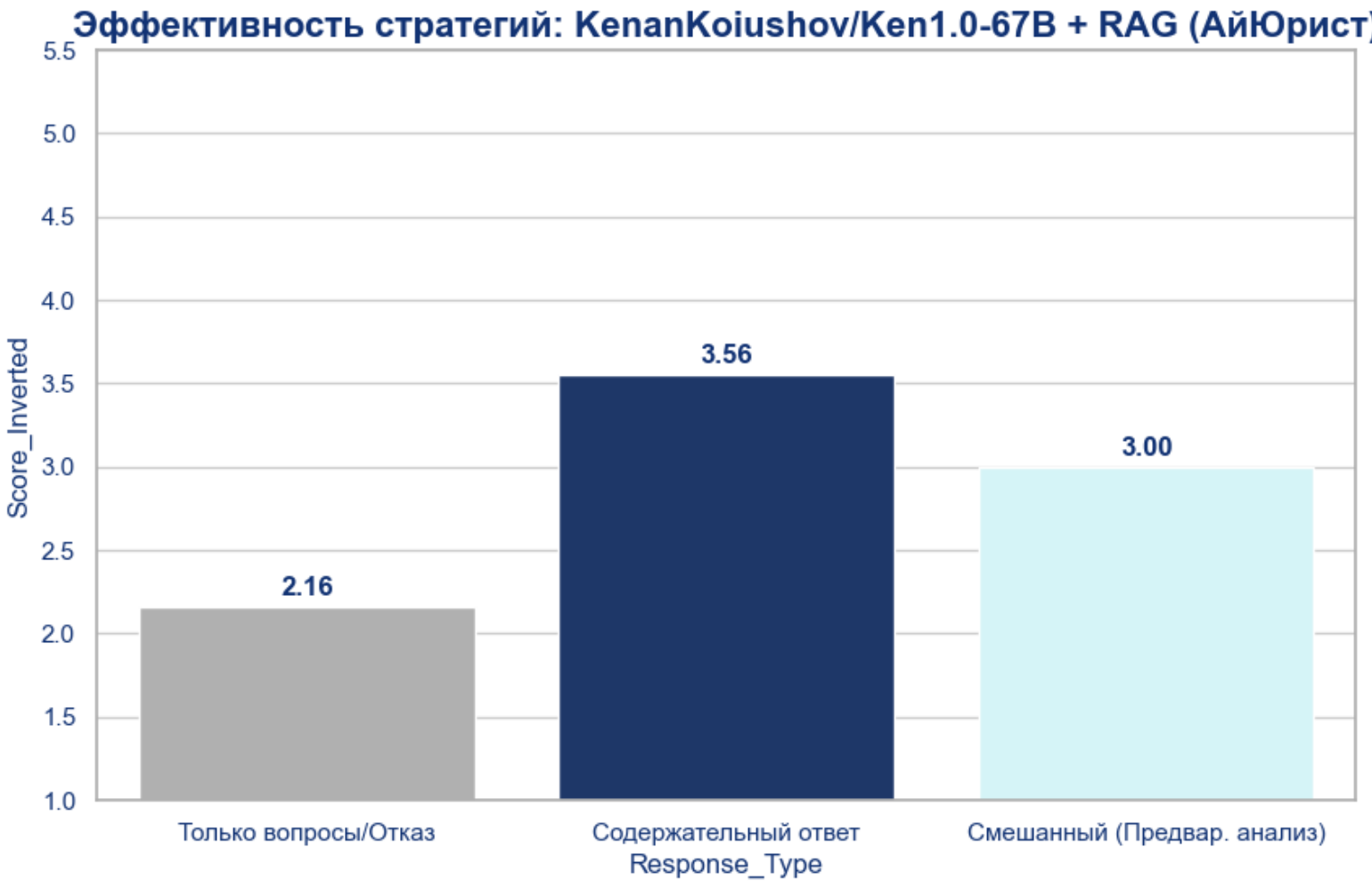
Эксперимент А: «Штраф за осторожность» (кейс Kenan+RAG)

Гипотеза: модель Kenan+RAG заняла последнее место не из-за низкого качества ответов, а *из-за стратегии поведения*.

При анализе текстов ответов выяснилось, что в 6 из 10 случаев (60%) модель отказалась давать прямой ответ, запрашивая уточнения: «Предоставьте документы», «Уточните обстоятельства», «Данных недостаточно».

Методика проверки: все оценки Kenan+RAG поделены на две группы:

- ✦ Группа А: Содержательные ответы (Q1, Q3, Q6) — модель дала решение задачи
- ✦ Группа В: Уточняющие вопросы (Q2, Q4, Q5, Q7, Q8, Q10) — модель ушла в отказ



Тип ответа	Средний балл	Количество оценок
Содержательный ответ	3.56	9
Только вопросы/отказ	2.16	19
Разница	**1.40**	

Разрыв в 1.4 балла — колоссальный. Когда модель даёт ответ, её качество (*3.56*) сопоставимо с лидерами рынка (*DeepSeek-V3: 3.61*). Провал в общем рейтинге вызван исключительно «штрафом» за уточняющие вопросы.

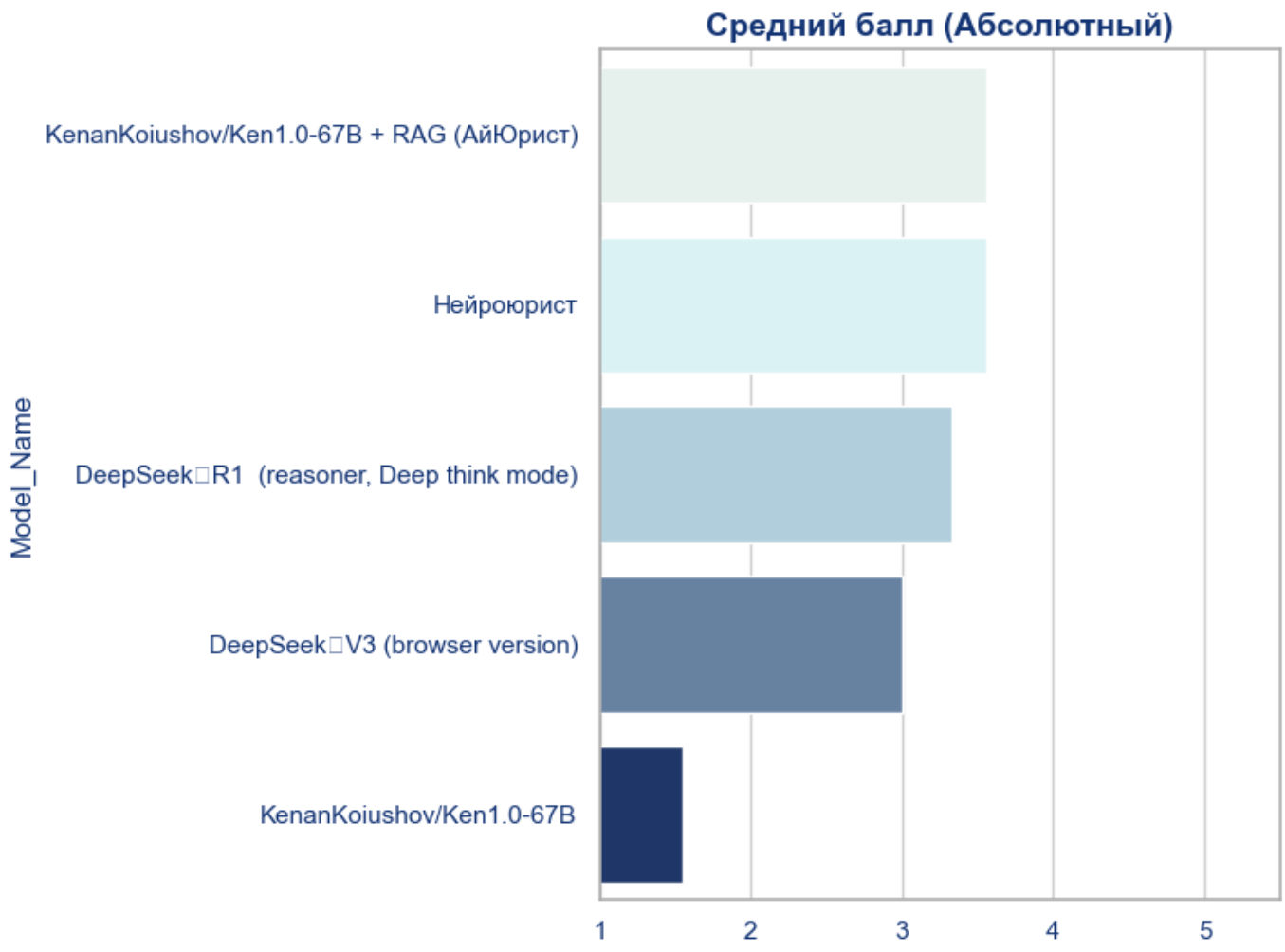
Причина поведения: судя по всему, у сервиса строгий промпт-инжиниринг, настраивающий модель на добывание дополнительной информации у пользователя. Это неплохо само по себе и действительно может привести в итоге к максимально качественному ответу. При этом в рамках бенчмарка другие модели выдают ответы сразу, и в ряде пользовательских сценариев получение моментального ответа — принципиально важно.

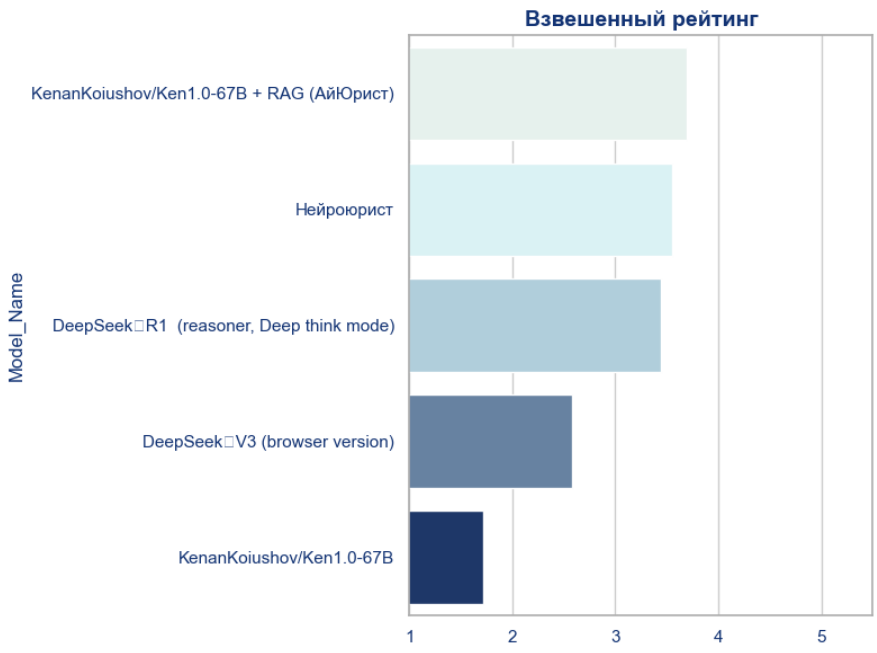
Вывод: у модели Kenan+RAG хорошие перспективы конкурировать с «высшей лигой» за 1-2 место либо при другом дизайне эксперимента, либо при корректировке системного промпта. При этом всего 4 вопросов и 9 оценок недостаточно для окончательных оптимистичных выводов.

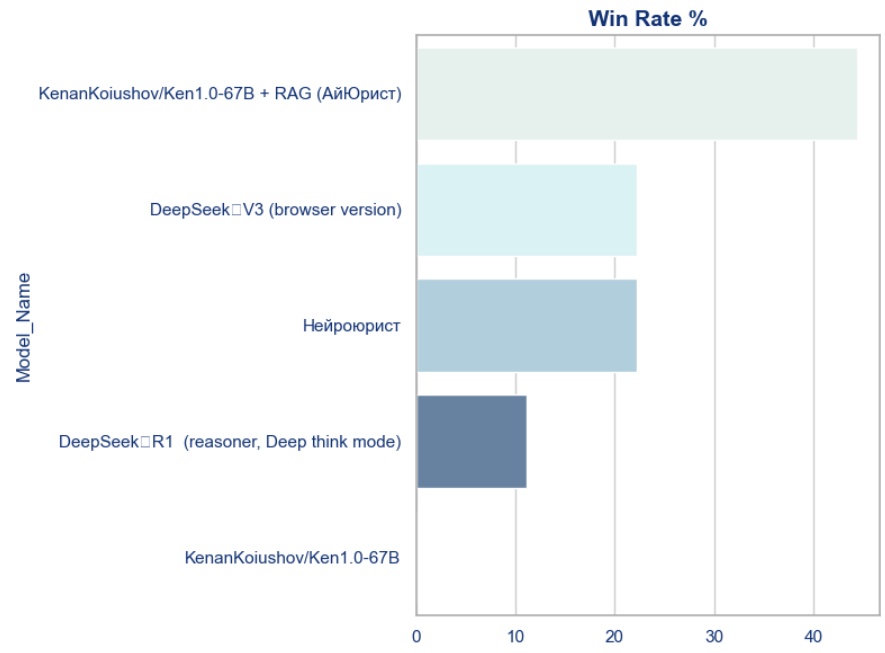
Эксперимент B: «Зелёные зоны» специализированных моделей

Гипотеза: Модели с RAG нужно оценивать в тех темах, где у них есть база знаний.

«Зелёная зона» Kenan+RAG (Q1, Q3, Q6):
Это вопросы, где модель не ушла в отказ и дала содержательный ответ.





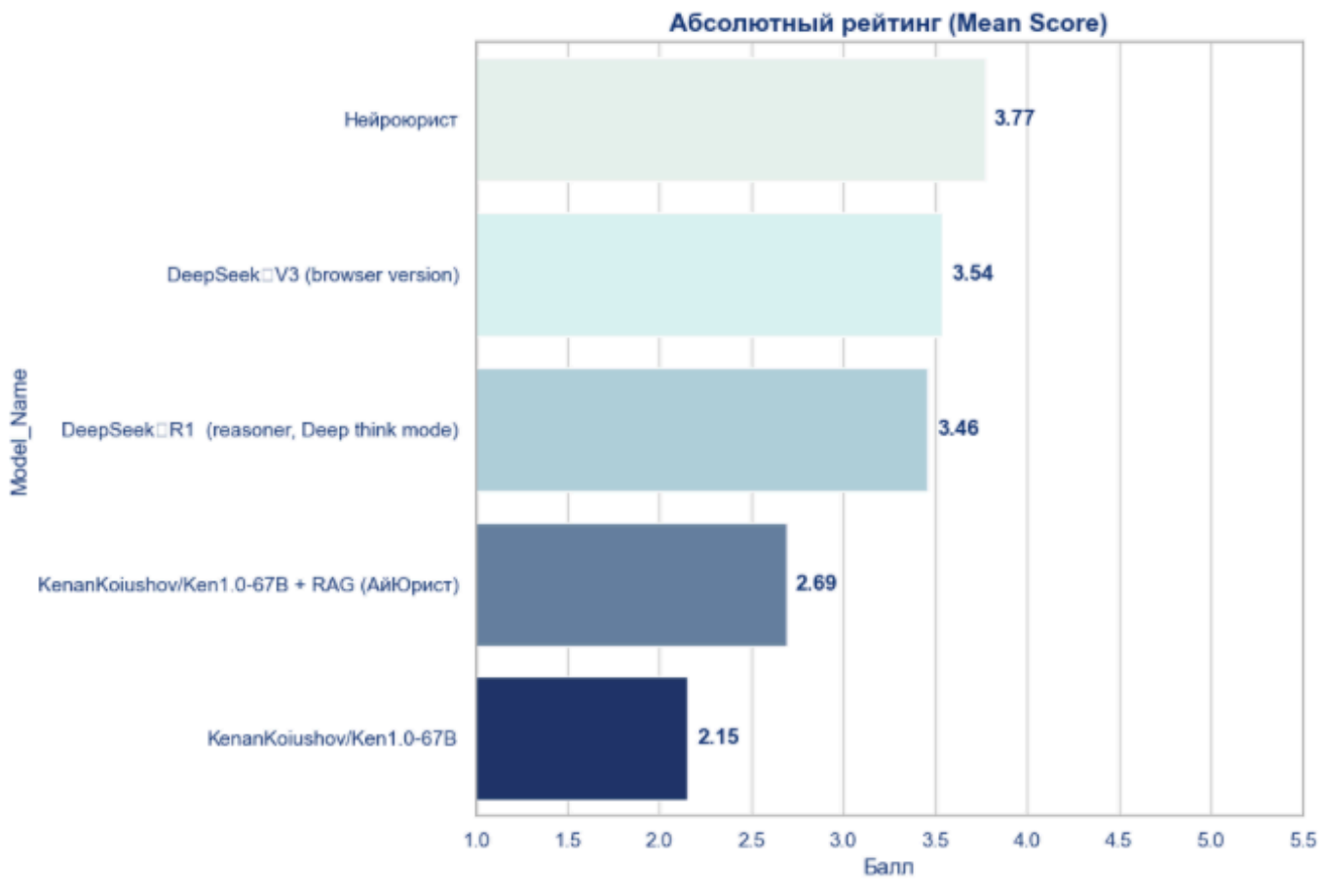


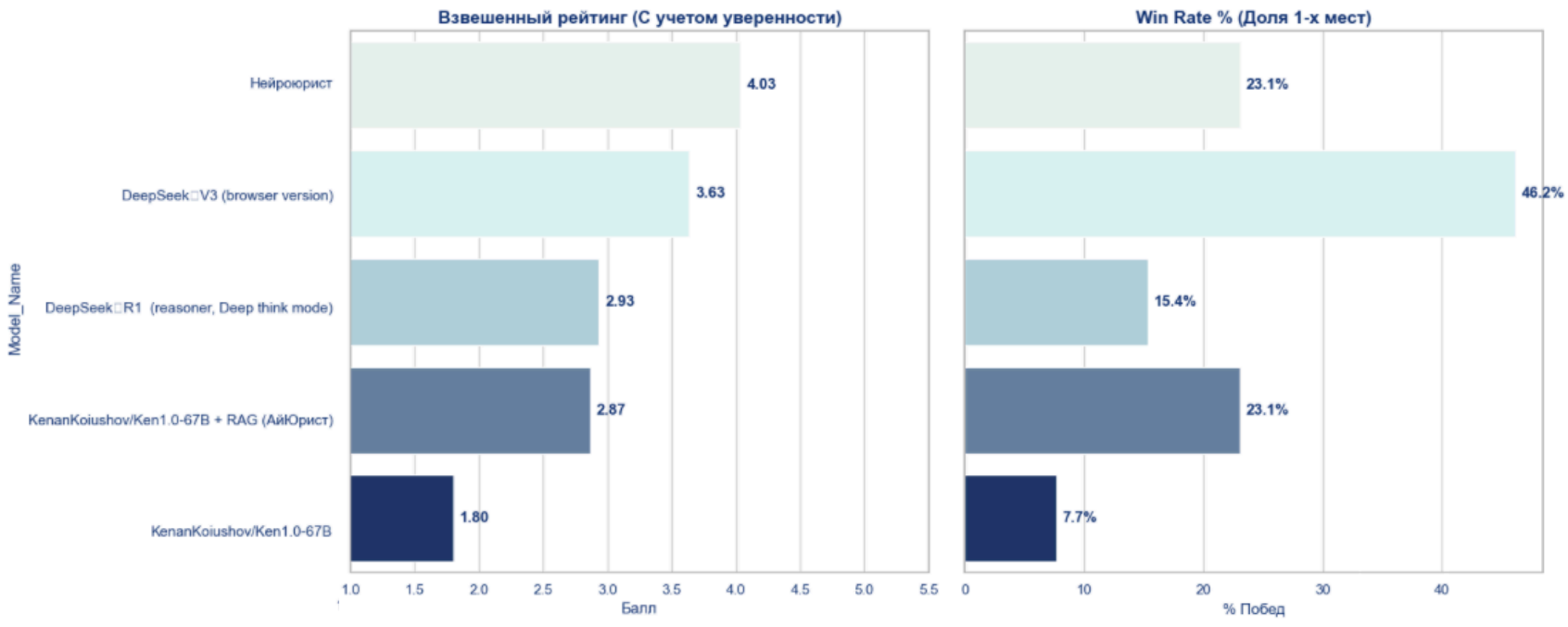
Модель	Взвешенный балл	Win Rate
Kenan+RAG	**3.69**	**44.4%**
Нейроюрист	3.55	
DeepSeek-R1	3.45	

В своей «зелёной зоне» Kenan+RAG занимает 1-е место, обгоняя всех конкурентов. Win Rate 44.4% означает, что почти в половине случаев ответ признаётся лучшим.

«Зелёная зона» Нейроюриста (Q3, Q4, Q7, Q8):

Это вопросы, соответствующие профилю базы знаний сервиса: обязательственное право, защита прав потребителей, рекламное право.





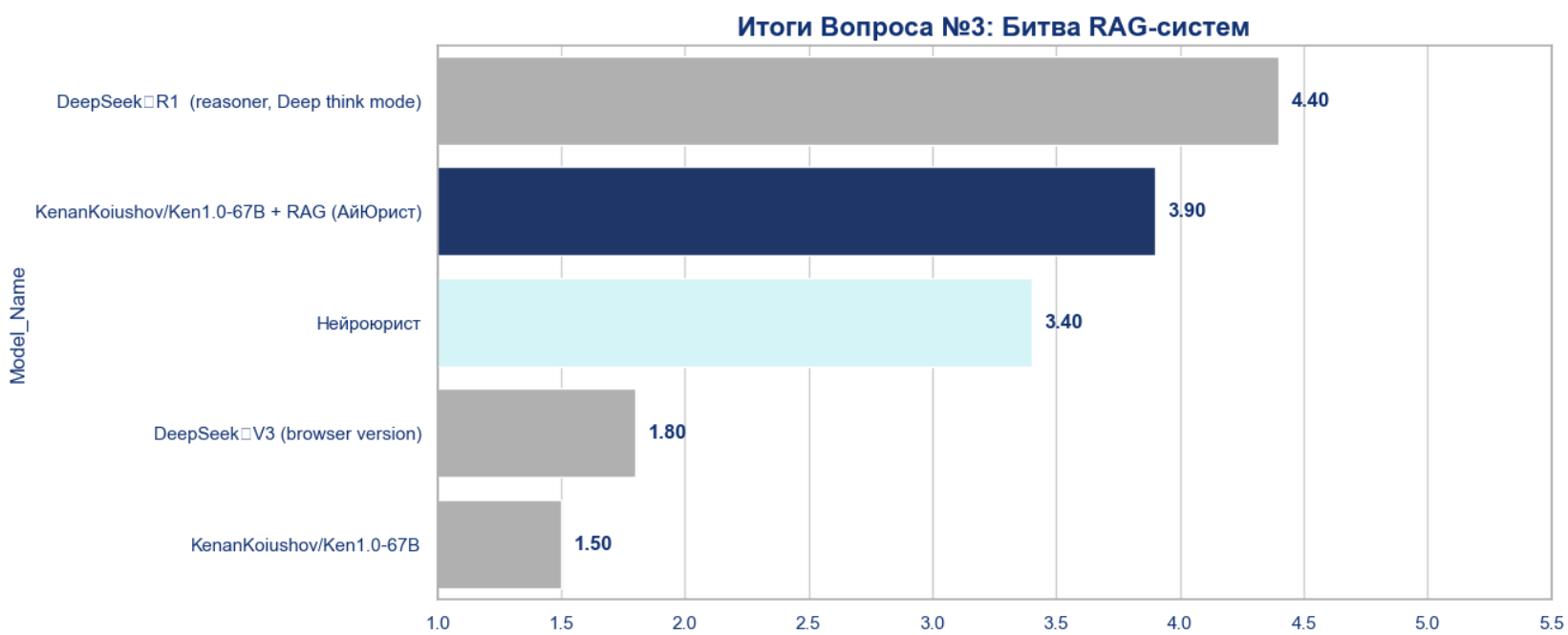
Модель	Взвешенный балл	Win Rate
Нейроюрист	**4.03**	**23%**
DeepSeek-R1	3.31	
DeepSeek-V3	2.93	

Рекордный результат: *4.03* балла — это абсолютный рекорд всего исследования. Ни одна модель ни в одном срезе не пробивала потолок в 4.0.

Нестабильность DeepSeek-V3: В этом же срезе модель имеет высокий Win Rate (*46%*), но провальный взвешенный балл (*2.93*). Эксперты-профи наказывают её за ошибки, которые не замечают неспециалисты.

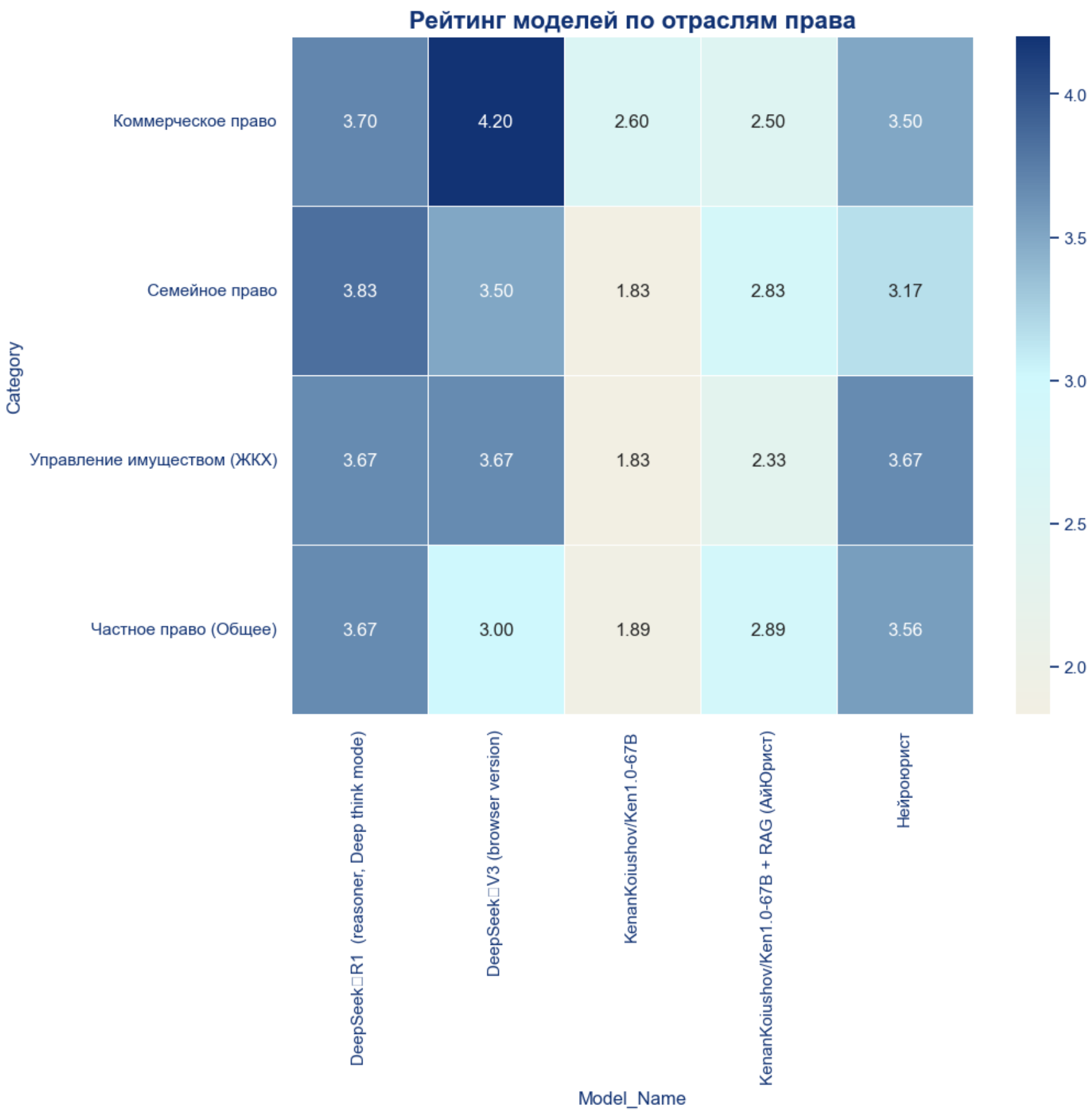
Эксперимент С: Сражение на «перекрёстке» (Q3)

Q3 (обеспечительный платеж) входит и в «зелёную зону» *Kenan+RAG*, и в профильную зону *Нейроюриста*. Это позволяет провести прямое сравнение двух RAG-систем.



Хотя обе модели показали высокие и сопоставимые результаты, эксперты всё равно отдали предпочтение «думающему» DeepSeek.

3.5. Анализ по отраслям права



Семейное право (Q1, Q10): *DeepSeek-R1*

Общие вопросы обязательственного права (Q2, Q3, Q4): *Нейроюрист* (в профильных темах) / *DeepSeek-R1* (в вещном праве)

Управление общедомовым имуществом (Q5, Q6): *Kenan+RAG*

Коммерческое право и специфика (Q7, Q8, Q9): *Нейроюрист*

РАЗДЕЛ 4

Итоги и выводы

4.1. Executive Summary

Исследование достигло поставленных целей и принесло практически значимые результаты.

Главный вывод: универсального лидера среди выбранных для сравнения моделей нет. Эффективность модели радикально зависит от типа задачи и наличия доступа к базе знаний.

Победитель общего зачёта — *DeepSeek-R1 (reasoner)*. Модель показала лучший баланс между глубиной анализа и стабильностью. Эксперты с высокой уверенностью отдают предпочтение ей.

Однако в специализированных задачах лидеры меняются. RAG-системы (*Нейроюрист* и *Kenan+RAG*) превосходят универсальные модели при условии попадания в профильную тему.

4.2. Детальная характеристика LLM-участников

🥇 *DeepSeek-R1* (reasoner) — надёжный и стабильный

Лидер общего зачёта (средний балл 3.71)

Сильные стороны:

- ✦ Самая высокая стабильность — минимальный разброс оценок
- ✦ Рост рейтинга при взвешивании на уверенность экспертов
- ✦ Хорошо справляется со сложными теоретическими конструкциями
- ✦ Меньше «галлюцинирует» по сравнению с V3

Слабые стороны:

- ✦ Уступает специализированным RAG-системам в профильных темах

Итог: лучший выбор для задач, требующих глубокого логического рассуждения, когда нет возможности использовать специализированный инструмент.

🥈 *Нейроюрист* — узкопрофильный специалист

3-е место в общем зачёте, 1-е место в профильных темах

Сильные стороны:

- ✦ Рекордный результат 4.03 балла в своих специализациях
- ✦ Интегрированная база актуального законодательства
- ✦ Стабильно высокие оценки от профильных экспертов

Слабые стороны:

- ✦ В непрофильных темах (семейное право, ЖКХ) работает как обычная не самая сильная LLM
- ✦ Ограниченное покрытие отраслей права

Итог: оптимальный выбор для задач по отраслям права, к которым заявлено наличие базы знаний.

🥉 *DeepSeek-V3 — красноречивый и убедительный*

2-е место в общем зачёте, лидер по Win Rate (32.3%)

Сильные стороны:

✦ «Вау-эффект» — часто занимает первые места

✦ Качественная структура и стиль изложения

✦ Широкий охват тем

Слабые стороны:

✦ Падение рейтинга при проверке экспертами с высокой уверенностью

✦ Склонность к убедительным «галлюцинациям»

Итог: хороший инструмент для генерации черновиков и простых консультаций, но требует обязательной проверки фактов специалистом.

Kenan+RAG (АйЮрист) — «спящий гигант»

Последнее место в общем зачёте, 1-е место в «зелёной зоне»

Сильные стороны:

✦ Качество ответов (когда даёт) на уровне лидеров рынка (*3.56*)

✦ Высокий Win Rate в профильных темах (*44.4%*)

✦ Принципиальный отказ от «галлюцинаций»

Слабые стороны:

✦ Избыточная осторожность — отказ отвечать в 60% случаев

✦ Жёсткий порог срабатывания RAG

Итог: в текущей настройке — модель с огромным потенциалом, но для проверки её эффективности необходим редизайн эксперимента или изменение системных настроек, чтобы модель давала результат с первого промпта.

Kenan (Base) — «стартовая точка»

Базовая модель без дополнительных настроек

Итог: в сравнении с версией +RAG демонстрирует, что Machine Learning показывает себя менее эффективным по сравнению с RAG. Для эксплуатации в текущем виде нужно хорошо продумать use cases.

4.3. Методологические выводы

1. RAG не мёртв. А очень важен для юридических задач.

Эксперимент наглядно доказал: модели с доступом к актуальной базе знаний превосходят не только «голые» LLM, но и дообученные методами машинного обучения. Даже модель среднего размера (*Kenan-67B*) с хорошо настроенным RAG способна конкурировать с топовыми универсальными системами.

2. Взвешенная оценка раскрывает истинную картину.

Простое среднее арифметическое искажает результаты в пользу «красноречивых» моделей. Учёт уверенности экспертов позволяет отфильтровать шум и выявить реальных лидеров по качеству.

3. Универсального супер-решения в выборке нет.

Наталкивает на мысль о том, что будущее юридического ИИ — не один «суперчатбот», а экосистема специализированных агентов под разные отрасли права и типы задач.

4. Для human-eval бенчмарка юридических задач необходимо проводить отдельную обстоятельную работу по подбору вопросов, чтобы избежать ситуаций низкой уверенности экспертов и противоречивых подходов при решении одной и той же задачи.

4.4. (Бес)полезность исследования

Можно ли доверять результатам? Мнение Gemini и Claude:

1. Если бы бенчмарк оказался неинформативен, все модели получили бы схожие баллы. Мы видим статистически значимую разницу между лидерами и аутсайдерами.

2. Метрика взвешенного рейтинга отфильтровала шум. Изменение позиций моделей при взвешивании — сигнал валидности методологии.

3. Бенчмарк показал не просто «кто лучше», а качественно разные типы поведения моделей (стабильный, красноречивый, осторожный).

Ограничения, влияющие на интерпретацию:

1. Малая выборка экспертов — результаты по отдельным вопросам могут быть смещены.

2. Низкое согласие по Q2, Q6, Q8 — эти вопросы отражают скорее спорность темы, чем качество моделей.

3. Формат теста (1 промпт = 1 ответ) не учитывает диалоговые возможности моделей.

Приложение 1

Глоссарий

Human evaluation benchmark (expert-based benchmark) — это бенчмарк, в котором качество ответов моделей на стандартизированный набор задач оценивается независимыми экспертами по заранее заданным критериям и шкалам, что позволяет количественно сравнивать модели между собой.

Средний балл (Mean Score) — сумма всех оценок, делённая на их количество. Показывает «типичный» результат модели.

Взвешенный рейтинг (Weighted Score) — средний балл, где каждая оценка умножена на коэффициент уверенности эксперта. Позволяет учесть, что мнение специалиста в своей области ценнее мнения «дженералиста». Формула расчёта: **`WeightedAvg = Σ(Оценка × Уверенность) / Σ(Уверенность)`**

Win Rate (процент побед) — доля случаев, когда модель заняла 1-е место. Показывает способность выдавать лучший результат.

Стандартное отклонение (Standard Deviation) — мера разброса оценок. Низкое значение = оценки кучные, модель стабильна. Высокое значение = большой разброс, модель непредсказуема.

Коэффициент Кендалла (Kendall's W) — показатель согласованности экспертов при ранжировании. Значения от 0 (полное несогласие) до 1 (полное единодушие). W > 0.7 считается высоким согласием. Формула расчёта: $W = 12S / (m^2(n^3 - n))$, где S — сумма квадратов отклонений сумм рангов от среднего, m — число экспертов, n — число объектов.

Корреляция — мера связи между двумя показателями. Положительная корреляция: когда один растёт, другой тоже растёт. r ≈ 0.4 — средняя связь.

RAG (Retrieval-Augmented Generation) — технология, при которой языковая модель дополнена системой поиска по базе знаний. Перед генерацией ответа модель ищет релевантную информацию в базе данных (например, в текстах законов или судебных решений).

Приложение 2

Материалы бенчмарка

Для обеспечения полной прозрачности исследования и для того, чтобы любой желающий мог перепроверить результаты, все основные материалы бенчмарка выложены в открытый доступ по **этой ссылке**.

Что можно найти на этом диске:

1. Файл benchmark_data.csv — карта всех 10 вопросов и ответов от всех 5 LLM. Формат .csv можно прочитать в Excel, а также удобно использовать для работы с любыми LLM или скриптами.
2. Папка responses с 11 файлами в формате .xlsx — это заполненные экспертами формами с их баллами и оригинальными комментариями. Организатором не вносилось каких-либо изменений в эти файлы.
3. Файл parsed_data.xlsx — «плоская» таблица, из которой вынесены все оценки от всех экспертов по всем вопросам, инвертированы для удобства анализа, раскрыты составившие ответы нейросети.
4. Файл анализ результатов.ipynb — это ноутбук с кодом, который обрабатывал данные экспертов, вынесенные в плоскую таблицу. Этот файл можно просмотреть в интерактивных блокнотах типа Jupyter Lab или Google Colab.
5. Файлы VISUAL_REPORT_by questions в формате .xlsx и .pdf — это анализ каждого отдельного вопроса, посчитанные для каждого вопроса ключевые метрики, сводные таблицы по оценкам всех экспертов с указанием их уровня уверенности в ответах.

Благодарности

Благодарю экспертов-оценщиков за их энтузиазм, экспертизу и совершенно бесценное время!

Дарья Бондарчук
Соорганизатор *Moscow Legal Hackers, а*втор Telegram-канала Юристы в кулуарах

Ольга Каменская
Адвокат по семейным и наследственным делам, советник МОКА по цифровизации. Медиатор Центра медиации при РСПП. Эксперт АНО «Метод». Автор Telegram-канала Слепая Фемида

Александр Клепцин
Начальник юридического отдела АО «Уралмостострой», ведёт хороший Telegram-канал о хорошем кино

Кирилл Кональт
Юрист-инхаус

Степан Леонтьев
Вольный legaltech стрелок

Сергей Русанов
Частнопрактикующий судебный юрист, автор Telegram-канала Споры о праве

Алексей Суворин
Юрист-энтузиаст ИИ

Алексей Стёпин
DPO в IT-компании

Михаил Тевс
Сооснователь сообщества *Нейросети | ilovedocs*

А также пожелавших остаться инкогнито **Оценщика 1** и **Оценщика 6**!
Что бы я без вас делала…

И, конечно, сообществу Нейросети | ilovedocs большое спасибо за информационную поддержку и среду, в которой возможны такие затеи!

Дисклеймеры

Использованные в материале изображения сгенерированы нейросетью NanoBanana Pro (что особенно заметно по рукам атлетов), концепт придуман автором.

При подготовке материала активно использовались нейросети: для кодинга форм для экспертов, их обработки, кода анализа результатов и построения графиков, анализа и интерпретации графиков, сборки текста отчёта. В таких задачах они, конечно, раскрываются на полную мощь и позволяют получить быстрый классный результат.

Связаться с организатором

yakunenko.ekaterina@gmail.com
https://t.me/delay_RAG