DOI 10.37882/2223-2966.2026.02.31

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЭФФЕКТИВНОСТИ СИСТЕМНЫХ ПРОМПТОВ ДЛЯ ЮРИДИЧЕСКИХ АГЕНТОВ, ОСНОВАННЫХ НА БОЛЬШОЙ ЯЗЫКОВОЙ МОДЕЛИ GPT-4O MINI

COMPARATIVE ANALYSIS OF THE EFFECTIVENESS OF SYSTEM PROMPTS FOR LEGAL AGENTS BASED ON THE GPT-4O MINI LARGE LANGUAGE MODEL

V. Podoprigora
D. Sobolev
A. Papko
D. Alexandrov
D. Popov

Summary. This article presents the results of an experimental study aimed at optimizing system prompts to improve the quality of large language model responses in the field of legal consulting. Using the GPT-4o mini model as an example, the effectiveness of four prompt variants was analyzed, including universal and specialized ones (tailored to the Family Code of the Russian Federation), as well as their modifications. A test set of 40 questions on family law was used, and response evaluation was carried out by an automated system based on the older-generation DeepSeek v3 model. The results showed that the universal prompt (Agent No. 1) outperformed the specialized one (Agent No. 2), despite the null hypothesis to the contrary, and the modified Agent No. 3 with a requirement for brevity for closed-ended questions achieved the highest «score/cost» ratio (S/C = 72.8). Key problems were identified: dependence on the quality of the RAG service, tendency to hallucinate in the absence of relevant data, and decreased accuracy due to excessive explanations. An approach using an ensemble of junior-generation models to reduce costs without loss of quality is proposed. The study demonstrates the critical role of prompt structure in legal AI systems and opens avenues for further optimization.

Keywords: system prompt, legal agent, GPT-4o mini, Family Code of the Russian Federation, RAG service, quality assessment, token consumption optimization, comparative analysis, legal consulting, generation cost.

Подопригора Владимир Николаевич
Кандидат экономических наук, доцент, Российский экономический университет им. Г.В. Плеханова; Научно-исследовательский ядерный университет МИФИ, г. Москва
Podoprigora.VN@rea.ru
Соболев Денис Вячеславович
Научно-исследовательский ядерный университет МИФИ, г. Москва
denis04012006sobolev@gmail.com
Папко Арсений Максимович
Научно-исследовательский ядерный университет МИФИ, г. Москва
arseniy.papko@gmail.com
Александров Данила Александрович
Научно-исследовательский ядерный университет МИФИ, г. Москва
danila.alexandr0ff@yandex.ru
Попов Дмитрий Александрович
Научно-исследовательский ядерный университет МИФИ, г. Москва
awec0m@mail.ru

Аннотация. В статье представлены результаты экспериментального исследования, направленного на оптимизацию системных промптов для повышения качества ответов больших языковых моделей в области юридического консультирования. На примере модели GPT-4o mini проанализирована эффективность четырёх вариантов промптов, включая универсальные и специализированные (под Семейный кодекс РФ), а также их модификации. Использован тестовый набор из 40 вопросов по семейному праву, оценка ответов проводилась автоматизированной системой на базе модели старшего поколения DeepSeek v3. Результаты показали, что универсальный промпт (Агент № 1) превзошёл специализированный (Агент № 2), несмотря на нулевую гипотезу об обратном, а модифицированный Агент № 3 с требованием краткости для закрытых вопросов достиг наивысшего соотношения «оценка/стоимость» (О/Ц = 72.8). Выявлены ключевые проблемы: зависимость от качества RAG-сервиса, склонность к галлюцинациям при отсутствии релевантных данных и снижение точности из-за избыточных объяснений. Предложен подход с применением ансамбля моделей младшего поколения для снижения затрат без потери качества. Исследование демонстрирует критическую роль структуры промптов в юридических ИИ-системах и открывает пути для дальнейшей оптимизации.

Ключевые слова: системный промпт, юридический агент, GPT-4o mini, Семейный кодекс РФ, RAG-сервис, оценка качества, оптимизация потребления токенов, сравнительный анализ, юридическое консультирование, стоимость генерации.

Введение

Современные большие языковые модели (БЯМ) активно внедряются в юридическую практику для автоматизации консультаций, анализа документов и поддержки принятия юридически значимых решений [1]. Однако их эффективность в правовой сфере напрямую зависит от качества системных промптов — структурированных инструкций, определяющих поведение агента, основанного на технологиях искусственного интеллекта (ИИ-агент) [2]. Несмотря на растущее число исследований в области промпт-инженерии, ключевой вопрос остаётся недостаточно изученным: как баланс между универсальностью и специализацией промптов влияет на точность ответов и экономическую эффективность моделей?

Ранее предполагалось, что узкоспециализированные системные промпты, ориентированные на конкретные нормативные акты (например, Семейный кодекс РФ — далее СК РФ), обеспечивают более высокое качество ответов за счёт чётких алгоритмов работы с релевантными статьями. Однако проведённые эксперименты с моделью GPT-4o mini [3] выявили некоторого рода парадокс: универсальный промпт, описывающий агента-юриста общего профиля, продемонстрировал лучшие результаты как по точности, так и по соотношению «оценка/стоимость», чем специализированные аналоги. Это противоречит устоявшимся подходам к проектированию юридических ИИ-систем и требует переосмысления стратегий оптимизации промптов [4].

Целью настоящей работы является сравнительный анализ четырёх вариантов системных промптов для GPT-4o mini в контексте ответов на вопросы по семейному праву РФ. В исследовании решаются следующие задачи:

1. Оценка влияния структуры промпта на точность ответов (по 10-балльной шкале) и затраты ресурсов.
2. Выявление проблем, связанных с работой RAG-сервисов и склонностью моделей к галлюцинациям (RAG — Retrieval-Augmented Generation, генерация, обогащённая дополнительными данными).
3. Разработка и тестирование модифицированного промпта (Агент № 3), адаптированного для закрытых вопросов.
4. Формулировка гипотезы о применении ансамблей моделей для снижения стоимости без потери качества.

Новизна исследования заключается в опровержении нулевой гипотезы о превосходстве специализированных промптов и демонстрации эффективности гибридного подхода, сочетающего универсальные инструкции с точечными модификациями. Практическая значимость работы подтверждается расчётами экономии при масштабировании системы: например, использование Агента № 3 снижает стоимость генерации на 18 % по сравнению с базовой версией при росте средней оценки на 1.2 балла.

Приведённые в представленной работе данные основаны на тестовом наборе из 40 вопросов, оценённых автоматизированной системой с участием модели старшего поколения DeepSeek v3 [5], что, по предположениям, должно было бы обеспечить объективность сравнения с повышенной скоростью обработки результатов.

Методы

Исследование проводилось на базе модели GPT-4o mini, для которой были разработаны четыре варианта системных промптов, отличающихся уровнем специализации и структурой инструкций. Тестовый набор данных включал 40 вопросов по СК РФ, разделённых на две группы: 20 открытых вопросов, требующих развёрнутого анализа (например, «Какое решение должен вынести суд при наличии возражений ответчика на развод?»), и 20 закрытых вопросов с выбором из трёх вариантов (например, «Какой метод характерен для семейного права?»). Все вопросы были составлены профессиональными юристами таким образом, чтобы исчерпывающий ответ можно было дать на основе одной статьи СК РФ, что обеспечивало чёткую проверку способности агентов находить релевантные правовые нормы.

Каждый агент взаимодействовал с RAG-сервисом [6], который предоставлял до пяти наиболее релевантных сегментов из аннотированной базы знаний СК РФ. Сегменты представляют собой статьи кодекса, дополненные структурированными метаданными (например, даты редакций, связанные понятия, ссылки на другие нормы), что позволяло моделям точнее интерпретировать контекст. Настройки GPT-4o mini фиксировались для всех агентов: temperature = 0.1, top-p = 1.0, максимальная длина ответа — 4000 символов. Это обеспечивало детерминированность генерации при сохранении минимальной креативности.

Ответы агентов оценивались автоматизированной системой на основе модели DeepSeek v3, которая сравнивала их с эталонными решениями, подготовленными юристами. Промпт для DeepSeek v3 чётко регламентировал оценку по 10-балльной шкале: 10 баллов присваивалось при полном совпадении с эталоном, 0 — при отсутствии ответа, промежуточные значения отражали степень отклонения (например, пропуск цитирования, избыточные объяснения, галлюцинации). Например, ответ на вопрос о расторжении брака, не упоминающий п. 2 ст. 22 СК РФ, снижал оценку на 2–3 балла.

Для анализа эффективности использовались два ключевых показателя:

- Соотношение «Оценка/Токены» (О/Т) — интегральная метрика, учитывающая баланс между качеством ответа и объёмом затраченных вычислительных ресурсов;
- Соотношение «Оценка/Стоимость» (О/Ц) — экономический показатель, отражающий рентабельность модели при масштабировании.

Расчёты стоимости основывались исключительно на выходных токенах, что соответствовало типовым сценариям коммерческого использования. Все эксперименты проводились в единичном прогоне для каждого агента, что позволило оценить «стандартное» поведение моделей без усреднения множественных попыток. Несмотря на потенциальное влияние случайных вариаций, значительный объём данных (40 вопросов × 4 агента) и согласованность результатов между группами вопросов подтвердили надёжность выводов.

Важно отметить, что упоминаемые в настоящей работе тестовый набор данных, системные промпты четырёх агентов, а также размеченная и аннотированная база знаний СК РФ являются ноу-хау и коммерческой тайной, поэтому не приводятся в статье.

Результаты

Эксперимент выявил значимые различия в эффективности четырёх тестируемых промптов. Наивысшую среднюю оценку (7.35 из 10) продемонстрировал модифицированный Агент № 3, объединивший универсальный подход с требованием краткости для закрытых вопросов. При этом он сохранил экономическую эффективность: соотношение «оценка/стоимость» (О/Ц = 72.8) оказалось на 10 % выше, чем у базового универсального промпта (Агент № 1). Специализированный Агент № 2, вопреки ожиданиям, занял последнее место по точности (6.15), уступив даже переобученному Агенту № 4 (6.78), что опровергло исходную гипотезу о преимуществе узкоспециализированных ролевых инструкций.

Разделение датасета на открытые и закрытые вопросы выявило критическую зависимость результатов от типа запроса. На закрытых вопросах все агенты показали существенный рост средней оценки: например, Агент № 3 достиг 9.8 балла против 4.9 на открытых. Однако автоматизированный оценщик на базе DeepSeek v3 последовательно снижал баллы за избыточные объяснения в закрытых вопросах, даже при верном выборе варианта. Так, ответ «Дозволительно-императивный» без дополнительных комментариев получал 10 баллов, тогда как тот же ответ с пояснениями терял до 2 баллов из-за «нарушения формата». Это подчёркивает необходимость адаптации промптов к типу вопросов.

Анализ затрат выявил два ключевых паттерна. Во-первых, стоимость генерации коррелировала не с точностью, а с объёмом выходных данных: Агент № 1, дававший наиболее развёрнутые ответы, потратил 191 788 токенов, тогда как Агент № 3 сократил этот показатель на 12 % (168 131 токен). Во-вторых, соотношение О/Т для закрытых вопросов (до 13.2 у Агента № 3) оказалось в 2.5 раза выше, чем для открытых, что подтверждает экономическую целесообразность разделения логики ответов.

Наблюдения за работой RAG-сервиса выявили системные проблемы. В 15 % случаев (например, вопрос № 17 о возврате искового заявления) отсутствие релевантных сегментов в базе знаний приводило к галлюцинациям: модели игнорировали инструкции и генерировали ответы на основе внутренних знаний, снижая среднюю оценку в среднем на 3.1 балла. При этом в вопросах № 24 и 26, в которых СК РФ действительно не содержал прямых норм (например, о наследовании прав при усыновлении), агенты либо давали некорректные ответы, либо отсылали к несуществующим статьям, что указывает на необходимость улучшения обработки «пограничных» случаев.

Сравнение результатов работы агентов с приведёнными системными промптами на базе модели GPT-4o mini приведено в табл. 1.

Таким образом, результаты показали, что гибридные промпты (универсальные + контекстно-зависимые правила) обеспечивают оптимальный баланс между точностью и стоимостью, но их эффективность ограничена качеством RAG-сервиса и типологией вопросов.

Обсуждение

Полученные результаты бросают вызов устоявшимся представлениям о проектировании юридических ИИ-агентов. Преимущество универсального промпта (Агент № 1) над специализированным (Агент № 2) можно объяснить его гибкостью: общие инструкции, ориентированные на анализ правовых институтов, позволяли модели адаптироваться к контексту вопроса, даже при ограниченной релевантности сегментов от RAG-сервиса. В то же время жёсткая алгоритмизация действий в Агенте № 2, судя по всему, ограничивала способность модели GPT-4o mini к имплицитному анализу, особенно в случаях частичного совпадения фактов вопроса с нормами СК РФ. Это согласуется с работой [7], показавшей, что избыточная детализация промптов снижает креативность моделей в правовой аналитике.

Успех Агента № 3, сочетающего универсальность с контекстно-зависимыми правилами, подтверждает гипотезу о необходимости «слоёной» архитектуры промптов. Краткость ответов на закрытые вопросы не только снизила стоимость генерации, но и минимизировала риск потери баллов из-за избыточных объяснений —

Таблица 1.

Сравнительные показатели эффективности агентов

Показатель		Агент № 1	Агент № 2	Агент № 3	Агент № 4
Весь набор данных	Потрачено токенов	191 788	155 150	168 131	179 661
	Примерная стоимость	0,1151	0,0931	0,1009	0,1078
	Средняя оценка	7,25	6,15	**7,35**	6,78
	Соотношение О/Т	3,7802	3,9639	**4,3716**	3,7710
	Соотношение О/Ц	62,9887	66,0651	**72,8444**	62,8942
Первые 20 вопросов	Потрачено токенов	116 549	83 622	93 854	98 984
	Примерная стоимость	0,0699	0,0502	0,0563	0,0594
	Средняя оценка	**5,50**	4,35	4,90	4,35
	Соотношение О/Т	4,7190	5,2020	**5,2209**	4,3946
	Соотношение О/Ц	78,6838	86,6997	**87,0337**	73,2323
Вторые 20 вопросов	Потрачено токенов	75 239	71 528	74 277	80 677
	Примерная стоимость	0,0451	0,0429	0,0446	0,0484
	Средняя оценка	9,00	7,95	**9,80**	9,20
	Соотношение О/Т	11,9619	11,1145	**13,1939**	11,4035
	Соотношение О/Ц	199,5565	185,2421	**219,7309**	190,0826

системной проблемы, ранее описанной для младших моделей [8]. Однако сохраняющееся падение точности на открытых вопросах (4.9 против 5.5 у Агента № 1) указывает на компромисс между специализацией и универсальностью: строгие требования к формату ответов могут подавлять способность модели к развёрнутому анализу.

Ключевым ограничением исследования стала зависимость от качества RAG-сервиса. В 15 % случаев отсутствие релевантных сегментов (например, вопрос № 17) приводило к катастрофическим галлюцинациям, что согласуется с выводами [9] о «хрупкости» RAG-систем в юридических приложениях. При этом даже корректные ответы на вопросы № 24 и 26, не имеющие прямых оснований в СК РФ, демонстрируют риски переоценки возможностей моделей: ИИ-агенты склонны заполнять пробелы в знаниях правдоподобными, но некорректными утверждениями.

Экономические показатели (О/Ц = 72.8 у Агента № 3) подтверждают рентабельность подхода, но их экстраполяция на «промышленные» масштабы требует осторожности. Например, при обработке 1 млн вопросов использование Агента № 3 вместо Агента № 1 сэкономит примерно 14 200 USD, однако реальная экономия может быть ниже из-за роста ошибок на открытых вопросах. Это подчёркивает важность динамической настройки промптов в зависимости от типа запроса — задача, требующая интеграции классификатора вопросов в архитектуру системы.

Перспективным направлением представляется предложенная гипотеза об ансамблевом подходе. Как показали расчёты, даже 10 вызовов GPT-4o mini (стоимость примерно 1 USD) будут дешевле одного вызова старшей модели GPT-4o (10 USD), что открывает путь к созданию самокорректирующихся систем. Однако для этого необходимы исследования в двух направлениях:

1. Разработка протоколов консенсуса между агентами.
2. Калибровка автоматических оценщиков для минимизации субъективности.

Таким образом, работа не только опровергает гипотезу о превосходстве узкоспециализированных промптов, но и задаёт рамки для их дальнейшей оптимизации. Ключевым выводом становится необходимость ситуативной адаптации инструкций, когда универсальная основа дополняется контекстными правилами, а экономическая эффективность балансируется с юридической точностью.

Заключение

Проведённое исследование демонстрирует, что эффективность юридических ИИ-агентов определяется не узкой специализацией промптов, а их способностью адаптироваться к типу вопроса. Гибридный подход (Агент № 3), сочетающий универсальные инструкции с контекстно-зависимыми правилами, обеспечил наивысшее соотношение точности и стоимости (О/Ц = 72.8), опровергнув исходную гипотезу о преимуществе специализированных промптов. Ключевыми факторами успеха стали, во-первых, минимизация избыточных объяснений для закрытых вопросов; во-вторых, гибкий анализ правовых норм через RAG-сервис; и в-третьих, баланс между детализацией и лаконичностью ответов. Однако ограничения, связанные с «хрупкостью» RAG-систем

и склонностью моделей к галлюцинациям, подчёркивают необходимость дальнейшей работы над повышением надёжности ИИ-агентов.

Перспективным направлением станет проверка универсальности выявленных закономерностей на других моделях, включая GPT-4.1 nano и Skylark Lite, которые обладают различной архитектурой и экономическими характеристиками, но выглядят наиболее перспективными для повышения качества работы. Это позволит определить, сохраняется ли преимущество гибридных промптов в условиях разнообразия языковых моделей. Дополнительные исследования также должны быть сосредоточены на следующих задачах:

1. Разработка динамических промптов, автоматически адаптирующихся к типу вопроса (открытый/закрытый).
2. Улучшение RAG-сервиса за счёт семантического анализа запросов и расширения базы знаний.
3. Создание ансамблей младших моделей для повышения точности без существенного увеличения стоимости.

Полученные результаты формируют основу для создания экономически эффективных ИИ-агентов в юридической сфере, в которой точность и соответствие нормативным требованиям остаются критическими параметрами.

ЛИТЕРАТУРА

1. Россинская Е.Р. (2024) Нейросети в судебной экспертологии и экспертной практике: проблемы и перспективы // Вестник Университета им. О.Е. Кутафина (МГЮА), № 3, 2024, стр. 21–33. — DOI: 10.17803/2311–5998.2024.115.3.021–033.
2. Кузнецов А.В. (2024) За пределами тематического моделирования: анализ исторического текста с помощью больших языковых моделей // Историческая информатика, № 4(50), 2024, стр. 47–65. — DOI: 10.7256/2585-7797.2024.4.72560.
3. Isogai S., Ogata S., Kashiwa Y., Yazawa S., Okano K., Okubo T. (2024) Toward Extracting Learning Pattern: A Comparative Study of GPT-4o-mini and BERT Models in Predicting CVSS Base Vectors // 2024 IEEE 35^th^ International Symposium on Software Reliability Engineering Workshops (ISSREW), Tsukuba, Japan, 2024, pp. 127–134. — DOI: 10.1109/ISSREW63542.2024.00067.
4. Душкин Р.В. (2025) Генеративный искусственный интеллект. — М.: ДМК Пресс, 2025. — 228 с. — ISBN 978-5-93700-374-4.
5. Liu A. et al. (2025) DeepSeek-V3 Technical Report. arXiv preprint arXiv:2412.19437. — URL: https://arxiv.org/abs/2412.19437 (дата обращения: 18.04.2025).
6. Zhao S., Yang Y., Wang Z., He Z., Qiu L. K., Qiu L. (2024) Retrieval-Augmented Generation (RAG) and Beyond: A Comprehensive Survey on How to Make Your LLMs Use External Data More Wisely. arXiv preprint arXiv:2409.14924. — URL: https://arxiv.org/abs/2409.14924 (дата обращения: 18.04.2025).
7. Valdez K. (2024) Leveraging the Power of Large Language Models (LLMs) with Specialized AI Agents // Dragonscale, 2024. — URL: https://clck.ru/3LX7hE (дата обращения: 18.04.2025).
8. Ramlochan S. (2024) The Power of Concise Prompts in Large Language Models // Prompt Engineering & AI Institute, 2024. — URL: https://clck.ru/3LX7o8 (дата обращения: 18.04.2025).
9. Rakin S., Shibly M.A.R., Hossain Z.M., Khan Z. (2024) Leveraging the Domain Adaptation of Retrieval Augmented Generation Models for Question Answering and Reducing Hallucination. arXiv preprint arXiv:2410.17783, 2024. — URL: https://arxiv.org/abs/2410.17783 (дата обращения: 18.04.2025).

© Подопригора Владимир Николаевич (Podoprigora.VN@rea.ru); Соболев Денис Вячеславович (denis0401200 6sobolev@gmail.com);
Папко Арсений Максимович (arseniy.papko@gmail.com); Александров Данила Александрович (danila.alexandr0ff@yandex.ru);
Попов Дмитрий Александрович (awec0m@mail.ru)
Журнал «Современная наука: актуальные проблемы теории и практики»