Слияние юридических знаний и ИИ: генерация с дополнением поиска с использованием векторных хранилищ, графов знаний и иерархической неотрицательной матричной факторизации

Дата: 2025-02-27 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2502.20364

Рейтинг: 65
Адаптивность: 75

Ключевые выводы:

Исследование представляет систему SMART-SLIC, которая объединяет Retrieval Augmented Generation (RAG) с векторными хранилищами, графами знаний и иерархической неотрицательной матричной факторизацией (NMFk) для улучшения работы с юридическими документами. Основная цель - создать более точную и интерпретируемую систему для анализа юридических текстов, которая минимизирует галлюцинации LLM и улучшает извлечение информации из сложных юридических документов.

Объяснение метода:

Исследование предлагает ценные концепции для эффективного поиска и анализа информации в LLM: многоаспектный подход, понимание иерархии документов, выявление связей и проверка фактов. Хотя техническая реализация недоступна обычным пользователям, концептуальное понимание может значительно улучшить формулирование запросов и оценку ответов LLM.

Ключевые аспекты исследования: 1. **Интеграция трех технологий для правовой информации**: Исследование предлагает систему, объединяющую векторные хранилища (Vector Stores), графы знаний (Knowledge Graphs) и неотрицательную матричную факторизацию (NMFk) для улучшения поиска и анализа юридической информации.

Иерархическая декомпозиция юридических текстов: Метод NMFk применяется для автоматического выделения тем и кластеризации юридических документов разных типов (конституция, законы, судебные дела), создавая многоуровневую структуру для более точного поиска.

Построение графа знаний для юридических связей: Система создает граф знаний, который формализует связи между юридическими документами (например, цитирования прецедентов, связи между законами), что позволяет выполнять структурированную навигацию.

Retrieval-Augmented Generation (RAG): Применение подхода RAG для минимизации "галлюцинаций" языковых моделей путем предоставления им доступа к фактической информации из юридической базы данных.

Экспериментальная оценка: Сравнение эффективности системы с существующими языковыми моделями (GPT-4, Gemini, Nemotron) в задачах поиска и анализа юридической информации.

Дополнение:

Применимость методов в стандартном чате

Исследование использует дообучение и API для реализации полной системы, но многие концепции и подходы можно адаптировать для стандартного чата с LLM без дополнительных технических средств:

Многоаспектный поиск информации: Вместо технической интеграции VS, KG и NMFk можно использовать структурированные запросы к LLM, разделяя их на:• Семантический поиск (значение и контекст)

• Структурный поиск (иерархические отношения)

• Тематический поиск (группировка по темам)

Пример: "Сначала объясни общую концепцию X, затем укажи ее связи с концепциями Y и Z, и наконец, опиши, к каким тематическим областям она относится"

Иерархическая декомпозиция:

Техническая NMFk-декомпозиция недоступна, но можно запросить LLM структурировать информацию иерархически Пример: "Раздели тему X на основные подтемы, затем для каждой подтемы выдели 3-5 ключевых аспектов"

Имитация графа знаний:

Запрашивать связи между концепциями явным образом Пример: "Какие концепции связаны с X? Для каждой связи объясни тип отношения и силу связи"

RAG через многоступенчатые запросы:

Запрашивать сначала источники, затем анализ на их основе Пример: "Перечисли 3-5 авторитетных источников по теме X. Теперь, основываясь только на этих источниках, ответь на вопрос Y"

Чанкинг через последовательные запросы:

Разбивать сложные темы на управляемые части Пример: "Давай разберем документ X по частям. Сначала проанализируй введение..." **Ожидаемые результаты от применения этих концепций**: - Снижение количества "галлюцинаций" в ответах LLM - Более структурированные и систематические ответы - Улучшенная возможность отслеживания источников информации - Более глубокое понимание взаимосвязей между различными концепциями - Возможность работать со сложными документами через их декомпозицию

Эти адаптированные подходы не достигнут технической эффективности полной системы из исследования, но значительно улучшат качество взаимодействия с LLM в стандартном чате.

Анализ практической применимости: 1. **Интеграция трех технологий - Прямая применимость**: Средняя. Обычный пользователь не может самостоятельно реализовать такую интегрированную систему, но концепция использования нескольких подходов к поиску может быть применена при формулировании сложных запросов к LLM. - **Концептуальная ценность**: Высокая. Понимание того, что комбинирование семантического поиска, структурированных связей и тематического анализа дает лучшие результаты, может помочь пользователям формулировать более эффективные запросы. - **Потенциал для адаптации**: Средний. Пользователи могут адаптировать идею многоаспектного поиска, например, запрашивая у LLM сначала общую информацию, затем связанные прецеденты, а затем тематический анализ.

Иерархическая декомпозиция юридических текстов Прямая применимость: Низкая. Метод требует специальных алгоритмов и не может быть напрямую использован пользователями. **Концептуальная ценность**: Высокая. Понимание иерархической структуры юридических документов помогает пользователям строить более точные запросы, запрашивая информацию на разных уровнях детализации. **Потенциал для адаптации**: Средний. Пользователи могут адаптировать концепцию, запрашивая у LLM сначала общие темы документа, а затем углубляясь в конкретные подтемы.

Построение графа знаний для юридических связей

Прямая применимость: Низкая. Построение графа знаний требует специализированных инструментов и данных. **Концептуальная ценность**: Высокая. Понимание связей между документами позволяет пользователям запрашивать у LLM не только прямую информацию, но и связанные материалы (например, "какие еще прецеденты связаны с этим законом?"). **Потенциал для адаптации**: Средний. Пользователи могут имитировать функциональность графа знаний, запрашивая у

LLM выявление связей между различными документами.

Retrieval-Augmented Generation (RAG)

Прямая применимость: Средняя. Пользователи не могут напрямую реализовать RAG, но могут использовать подход "проверки фактов" с LLM. **Концептуальная ценность**: Очень высокая. Понимание того, что LLM могут "галлюцинировать" и нуждаются в фактической проверке, критически важно для всех пользователей. **Потенциал для адаптации**: Высокий. Пользователи могут адаптировать RAG, запрашивая у LLM сначала поиск релевантных источников, а затем анализ на их основе.

Экспериментальная оценка

Прямая применимость: Средняя. Сравнительный анализ помогает пользователям понять сильные и слабые стороны различных LLM в юридических задачах. **Концептуальная ценность**: Высокая. Понимание того, что разные модели имеют разную точность в разных типах запросов, помогает выбрать подходящую модель для конкретной задачи. **Потенциал для адаптации**: Средний. Пользователи могут адаптировать подход к оценке, проверяя ответы LLM несколькими способами для критически важных запросов. Сводная оценка полезности: Предварительная оценка: 62/100

Исследование представляет значительную ценность для широкой аудитории, хотя большая часть технической реализации недоступна для непосредственного применения обычными пользователями. Основная ценность заключается в концептуальном понимании:

Важности многоаспектного подхода к поиску информации (семантический, структурный, тематический) Понимания иерархической структуры сложных документов Осознания взаимосвязей между различными документами Необходимости фактической проверки информации, генерируемой LLM Понимания различий в возможностях разных LLM Контраргументы к оценке: - Оценка могла бы быть выше (70-75), поскольку исследование предлагает очень четкие концептуальные модели, которые пользователи могут применить при формулировании запросов к LLM, даже не имея доступа к технической реализации. - Оценка могла бы быть ниже (50-55), поскольку исследование сфокусировано на юридической области, что ограничивает его применимость для пользователей, не работающих с юридическими документами.

Учитывая эти контраргументы, я корректирую оценку до 65/100. Исследование имеет высокую полезность, особенно в концептуальном плане, но требует значительной адаптации для применения в повседневном использовании LLM широкой аудиторией.

Оценка полезности: 65/100

Причины такой оценки: 1. Исследование демонстрирует эффективность

комбинирования различных подходов к поиску информации, что может быть адаптировано пользователями 2. Предлагает концептуальное понимание структуры сложных документов и связей между ними 3. Подчеркивает важность фактической проверки информации в LLM 4. Ограничено специфической (юридической) областью 5. Техническая реализация недоступна для обычных пользователей без специальных навыков

Уверенность в оценке: Уверенность в оценке: очень сильная

Я очень уверен в своей оценке, поскольку тщательно проанализировал как технические аспекты исследования, так и их потенциальную ценность для пользователей разного уровня технической подготовки. Исследование имеет четко выраженные концептуальные ценности, которые могут быть адаптированы пользователями, хотя прямая техническая реализация недоступна. Оценка учитывает баланс между технической сложностью и концептуальной применимостью.

Оценка адаптивности: Оценка адаптивности: 75/100

Исследование демонстрирует высокий потенциал адаптивности по следующим причинам:

Концептуальная адаптивность: Основные принципы (многоаспектный поиск, иерархическая структура, связи между документами) могут быть адаптированы пользователями при формулировании запросов к LLM, даже без технической реализации.

Методологическая адаптивность: Подход к декомпозиции сложных документов на темы и подтемы может быть имитирован через серию последовательных запросов к LLM.

Функциональная адаптивность: Функциональность графа знаний может быть частично воспроизведена через запросы о связях между различными документами.

Проверка фактов: Концепция RAG может быть адаптирована пользователями через запросы на проверку информации с указанием конкретных источников.

Отраслевая переносимость: Хотя исследование фокусируется на юридической области, принципы применимы к другим областям со сложной документацией (медицина, финансы, техническая документация).

Ограничения адаптивности включают необходимость специализированных инструментов для полной реализации подхода и специфику юридической области, которая может не полностью переноситься в другие контексты.

|| <Оценка: 65> || <Объяснение: Исследование предлагает ценные концепции для эффективного поиска и анализа информации в LLM: многоаспектный подход, понимание иерархии документов, выявление связей и проверка фактов. Хотя

техническая реализация недоступна обычным пользователям, концептуальное понимание может значительно улучшить формулирование запросов и оценку ответов LLM.> || <Адаптивность: 75>

Prompt:

Применение исследования SMART-SLIC в промптах для GPT
Ключевые элементы исследования для использования в промптах

Исследование SMART-SLIC демонстрирует эффективность комбинирования нескольких технологий для улучшения работы с юридическими документами: 1. **Retrieval Augmented Generation (RAG)** 2. **Векторные хранилища** для семантического поиска 3. **Графы знаний** для представления связей между документами 4. **Иерархическая неотрицательная матричная факторизация (NMFk)** для выявления тем

Пример промпта для юридического анализа

[======]

Юридический анализ с применением методологии SMART-SLIC Я хочу, чтобы ты выступил в роли юридического аналитика, используя принципы системы SMART-SLIC.

Контекст задачи

Мне нужно проанализировать следующий юридический документ: [ВСТАВИТЬ ТЕКСТ ДОКУМЕНТА]

Инструкции по анализу

Сначала разбей документ на логические фрагменты (chunking), выделяя ключевые разделы. Определи основные тематические кластеры в тексте, как это делается в NMFk. Создай концептуальную схему связей между выявленными темами, имитируя граф знаний. При ответе на мои вопросы: Цитируй конкретные разделы документа Указывай точные ссылки на источники Выделяй связи между различными частями документа Признавай неопределенность, если информации недостаточно
Первый вопрос

[ВСТАВИТЬ ВОПРОС ПО ДОКУМЕНТУ] [======]

Объяснение эффективности такого подхода

Данный промпт использует ключевые элементы методологии SMART-SLIC:

Разбиение текста (chunking) — исследование показало, что это улучшает точность поиска, повышая MRR до 0.65 для судебных дел.

Тематическое моделирование — имитирует работу NMFk по выявлению латентных тем, что помогает структурировать сложные юридические тексты.

Связи между документами — просьба создать концептуальную схему связей имитирует функциональность графа знаний.

Точность и прослеживаемость — требование цитировать конкретные разделы снижает вероятность "галлюцинаций" LLM, что было одним из ключевых преимуществ SMART-SLIC.

Хотя GPT не имеет прямого доступа к базам данных Neo4j или Milvus, используемым в исследовании, правильно структурированный промпт может имитировать некоторые аспекты этой методологии, значительно повышая качество анализа юридических документов.