DeepRAG: Поэтапное мышление при извлечении для крупных языковых моделей

Дата: 2025-02-03 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2502.01142

Рейтинг: 65
Адаптивность: 75

Ключевые выводы:

Исследование представляет DeepRAG - новую структуру для улучшения способности больших языковых моделей (LLM) к рассуждению с помощью поиска информации. Основная цель - моделирование процесса поиска как марковского процесса принятия решений, что позволяет LLM стратегически и адаптивно определять, когда использовать внешние знания, а когда полагаться на параметрические знания. Результаты показывают, что DeepRAG улучшает точность ответов на 21,99% при одновременном повышении эффективности поиска.

Объяснение метода:

DeepRAG предлагает ценную методологию декомпозиции сложных вопросов на подзапросы и определения необходимости внешнего поиска. Хотя полная техническая реализация недоступна обычным пользователям, концептуальные принципы могут быть адаптированы для более эффективного взаимодействия с LLM через структурированные запросы и пошаговое рассуждение.

Ключевые аспекты исследования: 1. **DeepRAG: моделирование процесса как MDP** - исследование представляет подход, который моделирует процесс поиска и использования внешней информации как марковский процесс принятия решений (MDP), что позволяет динамически определять, когда требуется обращение к внешним источникам данных.

Двухкомпонентная структура системы - DeepRAG включает две ключевые составляющие: "retrieval narrative" (структурированный поток поисковых запросов) и "atomic decisions" (решения о необходимости поиска для каждого подзапроса), что обеспечивает стратегический и адаптивный подход к поиску.

Метод бинарного дерева поиска - для каждого подзапроса система строит бинарное дерево, исследуя два возможных пути: использование параметрических знаний модели или обращение к внешней базе знаний.

Двухэтапное обучение модели - сначала применяется имитационное обучение на

синтезированных данных, затем используется "chain of calibration" для улучшения понимания моделью своих границ знаний.

Значительное улучшение точности и эффективности - эксперименты показали повышение точности ответов на 21-99% при сокращении количества обращений к внешним источникам по сравнению с другими методами.

Дополнение:

Применимость методов в стандартном чате

Хотя в исследовании используется дообучение модели и специализированное API для реализации полной системы DeepRAG, многие концептуальные подходы могут быть адаптированы для использования в стандартном чате с LLM без технических модификаций:

Структурированная декомпозиция вопросов Пользователи могут вручную разбивать сложные вопросы на последовательность подзапросов Для каждого подзапроса можно получать промежуточный ответ перед переходом к следующему шагу

Осознанное использование внешней информации

Пользователи могут самостоятельно решать, когда запрашивать модель о поиске дополнительной информации Можно явно указывать модели, когда ответ должен основываться на её параметрических знаниях

Итеративное построение ответа

Использование промежуточных ответов как основы для формулирования следующих подзапросов Постепенное построение полного ответа на основе собранных промежуточных результатов Ожидаемые результаты от применения этих концепций: - Повышение точности ответов на сложные вопросы - Снижение вероятности галлюцинаций модели - Более структурированное и прозрачное рассуждение - Лучшее понимание пользователем процесса формирования ответа

Эти адаптированные подходы не требуют дообучения или специального API, но могут значительно улучшить качество взаимодействия с LLM в стандартном чате.

Анализ практической применимости: 1. **Моделирование процесса как MDP**: - Прямая применимость: Низкая для обычных пользователей, так как требует глубокого понимания марковских процессов и сложной реализации. - Концептуальная ценность: Высокая, поскольку демонстрирует важность структурированного подхода к декомпозиции сложных вопросов и принятия решений о поиске информации. - Потенциал для адаптации: Средний, концепция пошагового разбиения вопросов может быть адаптирована пользователями для более эффективного взаимодействия с LLM.

Двухкомпонентная структура системы: Прямая применимость: Средняя, пользователи могут частично имитировать этот подход, разбивая сложные вопросы на подзапросы. Концептуальная ценность: Высокая, помогает понять, как эффективнее формулировать запросы к LLM для получения точных ответов. Потенциал для адаптации: Высокий, структура "retrieval narrative" может быть упрощена до пошагового метода формулирования запросов.

Метод бинарного дерева поиска:

Прямая применимость: Низкая, требует технической реализации. Концептуальная ценность: Средняя, демонстрирует преимущества сравнения результатов с использованием внешних источников и без них. Потенциал для адаптации: Средний, пользователи могут вручную проверять, требуется ли дополнительная информация для ответа.

Двухэтапное обучение модели:

Прямая применимость: Низкая, требует специальных навыков дообучения моделей. Концептуальная ценность: Средняя, показывает важность калибровки моделей для понимания границ их знаний. Потенциал для адаптации: Низкий, сложно реализовать обычным пользователям.

Улучшение точности и эффективности:

Прямая применимость: Высокая, демонстрирует конкретные преимущества структурированного подхода к запросам. Концептуальная ценность: Высокая, подтверждает эффективность методологии. Потенциал для адаптации: Высокий, пользователи могут адаптировать принципы для повышения качества своих взаимодействий с LLM.

Prompt:

Применение DeepRAG в промптах для GPT ## Ключевые принципы DeepRAG

DeepRAG представляет структуру для улучшения способности языковых моделей к рассуждению с использованием внешней информации через: - Стратегическое определение, когда использовать внешние знания - Декомпозицию сложных запросов на подзапросы - Оптимизацию поисковых операций

Пример промпта с применением принципов DeepRAG

[======] # Запрос с применением DeepRAG-подхода

Контекст задачи Я исследую влияние изменения климата на миграцию видов в экосистемах коралловых рифов.

Структурированный подход (по DeepRAG) 1. Сначала определи, какие аспекты

этой темы ты уже знаешь достаточно хорошо, а для каких потребуется дополнительная информация. 2. Декомпозируй основной вопрос на следующие подвопросы: - Какие ключевые механизмы влияния изменения климата на коралловые рифы? - Какие виды наиболее чувствительны к этим изменениям? - Какие существуют паттерны миграции в ответ на эти изменения? 3. Для каждого подвопроса: - Сначала ответь на основе своих параметрических знаний - Четко обозначь, где твои знания могут быть неполными или устаревшими - Предложи, какие конкретные внешние источники могли бы дополнить твой ответ

Формат ответа - Используй дерево рассуждений, четко показывая связи между подвопросами - В финальном ответе синтезируй информацию из всех подвопросов - Укажи степень уверенности в различных частях ответа [======]

Как работают принципы DeepRAG в этом промпте

Структурированное повествование поиска: Промпт декомпозирует сложный запрос на конкретные подзапросы, что позволяет модели более целенаправленно использовать свои знания.

Атомарные решения: Модель должна явно определить, для каких аспектов она имеет достаточно знаний, а для каких требуется внешняя информация.

Бинарный поиск по дереву: Промпт побуждает модель исследовать различные пути рассуждения, выбирая оптимальные на основе имеющихся знаний.

Калибровка границ знаний: Требование указать степень уверенности и потенциальные пробелы в знаниях помогает модели лучше осознавать границы своих возможностей.

Такой подход позволяет получить более глубокие, структурированные и обоснованные ответы, с четким разграничением между параметрическими знаниями модели и областями, где требуется дополнительная информация.