AirRAG: Активация внутреннего размышления для генерации с дополнением извлечения с использованием поиска на основе деревьев

Дата: 2025-02-14 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2501.10053

Рейтинг: 68
Адаптивность: 75

Ключевые выводы:

Исследование представляет новый метод AirRAG (Activating Intrinsic Reasoning for Retrieval Augmented Generation), который улучшает способности больших языковых моделей (LLM) в задачах рассуждения с использованием поиска на основе дерева. Основная цель - преодолеть ограничения существующих методов RAG, которые часто ограничены единственным пространством решений при работе со сложными задачами. AirRAG значительно превосходит существующие итеративные или рекурсивные подходы RAG, активируя внутренние способности рассуждения LLM и расширяя пространство решений контролируемым образом.

Объяснение метода:

AirRAG предлагает ценные концепции для эффективного взаимодействия с LLM: декомпозицию сложных задач, пять действий рассуждения, переформулирование запросов и рассмотрение проблемы с разных точек зрения. Хотя MCTS недоступен обычным пользователям, основные принципы можно адаптировать в повседневном использовании чатов, что делает исследование полезным для широкой аудитории.

Ключевые аспекты исследования: 1. **AirRAG (Activating Intrinsic Reasoning for RAG)** - метод, который использует древовидный поиск (Monte Carlo Tree Search, MCTS) для активации внутренних рассуждений в моделях LLM и расширения пространства решений для сложных задач.

Пять фундаментальных действий рассуждения - авторы разработали пять основных действий: системный анализ (SAY), прямой ответ (DA), ответ через поиск (RA), трансформация запроса (QT) и суммирующий ответ (SA), которые эффективно решают различные типы запросов.

Самосогласованность и масштабирование вывода - AirRAG использует самосогласованность и MCTS для исследования потенциальных путей рассуждения

и эффективного масштабирования вычислений при выводе.

Вычислительно оптимальные стратегии - метод применяет больше вычислительных ресурсов к ключевым действиям, что повышает общую производительность.

Модульная архитектура - AirRAG имеет гибкую структуру, позволяющую легко интегрировать другие передовые методы.

Дополнение:

Применимость методов в стандартном чате без дообучения или API

Методы AirRAG действительно требуют специальной реализации и API для полноценного функционирования в том виде, в котором они представлены в исследовании (особенно Monte Carlo Tree Search). Однако многие концепции и подходы можно адаптировать для работы в стандартном чате.

Концепции для адаптации в стандартном чате:

Пять фундаментальных действий рассуждения: Системный анализ (SAY): Пользователи могут просить LLM сначала проанализировать проблему и разбить ее на подзадачи Прямой ответ (DA): Запрос на использование только внутренних знаний модели Ответ через поиск (RA): В стандартном чате можно реализовать как пошаговые уточняющие вопросы Трансформация запроса (QT): Пользователи могут просить модель переформулировать исходный запрос для лучших результатов Суммирующий ответ (SA): Запрос на обобщение информации из предыдущих шагов

Декомпозиция сложных задач:

Пользователи могут явно запрашивать разбиение сложной задачи на компоненты Можно использовать пошаговое решение с промежуточными вопросами

Самосогласованность:

Запрос на генерацию нескольких подходов к решению задачи Просьба проанализировать сильные и слабые стороны каждого подхода Запрос на синтез наиболее надежного решения на основе всех подходов ### Ожидаемые результаты от применения:

Повышение точности ответов: Особенно для сложных многоэтапных задач **Более структурированные ответы**: Лучшая организация информации **Более надежные решения**: За счет рассмотрения проблемы с разных сторон **Лучшее понимание процесса рассуждения**: Пользователи получают доступ к промежуточным шагам мышления модели Эти адаптированные подходы могут значительно улучшить взаимодействие с LLM в стандартном чате без необходимости в специальной технической реализации или API.

Анализ практической применимости: 1. **Пять фундаментальных действий рассуждения**: - Прямая применимость: Средняя. Пользователи не могут напрямую реализовать эти действия в стандартном чате, но могут адаптировать подход при формулировании сложных запросов. - Концептуальная ценность: Высокая. Понимание этих действий помогает пользователям осознать, как LLM решают сложные задачи и как структурировать свои запросы. - Потенциал для адаптации: Высокий. Пользователи могут адаптировать эти действия в свои запросы, например, сначала запрашивая анализ проблемы, затем задавая уточняющие вопросы.

Самосогласованность и масштабирование вывода: Прямая применимость: Низкая. Пользователи не могут напрямую контролировать число выводов модели. Концептуальная ценность: Высокая. Понимание того, что несколько подходов к одной задаче могут дать более точный ответ, ценно для пользователей. Потенциал для адаптации: Средний. Пользователи могут запрашивать модель рассмотреть проблему с разных сторон или предложить несколько решений.

Системный анализ и декомпозиция задач:

Прямая применимость: Высокая. Пользователи могут напрямую просить модель проанализировать и разбить сложную задачу на подзадачи. Концептуальная ценность: Очень высокая. Понимание важности декомпозиции сложных задач - ключевой инсайт для эффективного использования LLM. Потенциал для адаптации: Очень высокий. Пользователи могут легко адаптировать этот подход к своим запросам.

Трансформация запросов:

Прямая применимость: Средняя. Пользователи могут применять некоторые техники переформулирования запросов. Концептуальная ценность: Высокая. Понимание того, как формулировка запроса влияет на качество ответа, очень ценно. Потенциал для адаптации: Высокий. Пользователи могут научиться переформулировать свои запросы для получения лучших результатов.

Гибкая архитектура:

Прямая применимость: Низкая. Обычные пользователи не могут изменять архитектуру LLM. Концептуальная ценность: Средняя. Понимание модульности подходов к решению задач может быть полезно. Потенциал для адаптации: Средний. Пользователи могут комбинировать различные подходы при работе с LLM.

Prompt:

Использование знаний из исследования AirRAG в промптах для GPT Исследование AirRAG предлагает ценные стратегии для улучшения рассуждений в больших языковых моделях. Вот как можно применить эти знания в промптах.

Пример промпта, использующего принципы AirRAG

[======] Я хочу, чтобы ты решил следующую сложную задачу, используя структурированный подход на основе древовидного рассуждения:

[ОПИСАНИЕ ЗАДАЧИ]

Пожалуйста, действуй следующим образом:

СИСТЕМНЫЙ АНАЛИЗ: Сначала проанализируй структуру проблемы, разбей её на ключевые компоненты и определи, какая информация потребуется для решения.

ТРАНСФОРМАЦИЯ ЗАПРОСА: Сформулируй 2-3 альтернативных подхода к решению проблемы или переформулируй задачу разными способами, чтобы увидеть её под разными углами.

ПРЯМОЙ ОТВЕТ: Попробуй дать предварительный ответ на основе имеющейся информации.

ОТВЕТ НА ОСНОВЕ ПОИСКА: Укажи, какую дополнительную информацию было бы полезно найти, и как бы ты использовал эту информацию.

СУММИРУЮЩИЙ ОТВЕТ: Объедини результаты предыдущих шагов в окончательное решение, указав наиболее вероятный верный ответ и обоснование.

Для каждого шага рассмотри несколько возможных направлений мысли, а не только первый пришедший в голову вариант. [======]

Объяснение подхода

Этот промпт использует ключевые принципы AirRAG:

Пять фундаментальных действий рассуждения - промпт явно структурирует процесс рассуждения в соответствии с действиями, предложенными в исследовании.

Древовидное пространство рассуждений - запрос на рассмотрение нескольких возможных направлений мысли имитирует древовидный поиск, позволяя модели исследовать различные пути решения.

Приоритизация ключевых действий - особое внимание уделяется системному анализу и трансформации запроса, которые согласно исследованию требуют большего разнообразия.

Самосогласованность - суммирующий ответ позволяет модели интегрировать результаты различных путей рассуждения и выбрать наиболее согласованное решение.

Такой подход особенно эффективен для сложных многоэтапных задач, требующих глубокого рассуждения и интеграции различных источников информации.