Агентное извлечение информации

Дата: 2025-02-22 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2410.09713

Рейтинг: 75
Адаптивность: 85

Ключевые выводы:

Исследование представляет концепцию агентного информационного поиска (Agentic IR) как новой парадигмы, расширяющей традиционный информационный поиск с помощью LLM-управляемых ИИ-агентов. Основная цель - переопределить информационный поиск от статического получения элементов информации к достижению целевых информационных состояний в динамической среде.

Объяснение метода:

Исследование вводит концепцию агентного информационного поиска, переопределяя взаимодействие с LLM как достижение "информационного состояния", а не просто получение информации. Предлагает практичный подход к многошаговому взаимодействию с LLM для решения комплексных задач. Концепции и примеры применимы сразу, без дополнительных инструментов, хотя полная реализация некоторых возможностей может требовать API-доступа.

Ключевые аспекты исследования: 1. **Новая парадигма информационного поиска**: Исследование вводит концепцию "Agent IC Information Retrieval" (Агентного информационного поиска), который переопределяет информационный поиск от простого получения релевантных элементов из предопределенного корпуса к достижению желаемого "информационного состояния" в динамической среде.

Информационное состояние как объект поиска: Вместо статичных информационных элементов вводится понятие "информационного состояния" пользователя, которое включает не только полученную информацию, но и контекст, предпочтения пользователя и процессы принятия решений.

Архитектура агентных систем: Исследование описывает архитектуру агентных систем для информационного поиска, включая ключевые компоненты агентов (профиль, память, планирование, действия) и дизайн систем (одноагентные и мультиагентные).

Практические применения: Представлены два конкретных практических применения - персональный ассистент и бизнес-ассистент, демонстрирующие возможности агентного информационного поиска в реальных сценариях.

Оценка и оптимизация систем: Предложены новые метрики и протоколы оценки для агентных систем, а также методы их оптимизации, учитывающие не только релевантность результатов, но и эффективность, полезность и этические аспекты.

Дополнение: Исследование представляет концепцию агентного информационного поиска, которая может быть применена в стандартном чате без необходимости дообучения или API. Хотя авторы для своих экспериментов могли использовать расширенные возможности, основные концепции применимы в обычном взаимодействии с LLM.

Концепции и подходы для применения в стандартном чате:

Информационное состояние как цель: Вместо простого запроса информации, пользователь может сформулировать желаемое "информационное состояние" - конечный результат, которого он хочет достичь. Например: "Я хочу спланировать поездку в Японию на 7 дней с бюджетом $2000".

Многошаговое взаимодействие: Пользователь может разбить сложную задачу на последовательность шагов и провести модель через эти шаги. Например, сначала определить маршрут, затем жилье, затем активности.

Итеративное уточнение: Пользователь может постепенно уточнять информацию на основе промежуточных результатов. Например: "Теперь, когда мы выбрали города, давай подберем отели в каждом из них".

Планирование действий: Пользователь может явно попросить модель составить план действий для достижения цели. Например: "Составь план действий для подготовки научной статьи".

Проактивный сбор информации: Пользователь может попросить модель определить, какая дополнительная информация нужна для решения задачи. Например: "Какую еще информацию тебе необходимо знать, чтобы помочь мне выбрать оптимальный маршрут?"

Ожидаемые результаты применения:

Более структурированные и целенаправленные взаимодействия с LLM Повышение эффективности решения сложных задач Улучшение качества получаемой информации благодаря более четкому определению цели Более персонализированные результаты из-за постепенного уточнения предпочтений Снижение когнитивной нагрузки на пользователя при решении сложных задач Даже без дополнительных API или инструментов, применение этих концепций может значительно улучшить опыт взаимодействия с LLM и повысить качество получаемых результатов.

Анализ практической применимости: 1. **Новая парадигма информационного поиска**: - Прямая применимость: Концепция агентного поиска помогает

пользователям переосмыслить взаимодействие с LLM, перейдя от простого поиска информации к достижению желаемого результата через последовательность действий. - Концептуальная ценность: Высокая. Пользователи получают понимание, что взаимодействие с LLM может быть многошаговым процессом, направленным на достижение конкретной цели. - Потенциал для адаптации: Концепция легко адаптируется к повседневным запросам, позволяя пользователям формулировать более комплексные задачи для LLM.

Информационное состояние как объект поиска: Прямая применимость: Пользователи могут переформулировать запросы, фокусируясь не на поиске информации, а на достижении конкретного "состояния" (например, получить не просто информацию о ресторанах, а забронировать столик). Концептуальная ценность: Очень высокая. Помогает понять, как формулировать запросы для получения не просто информации, а конкретного результата. Потенциал для адаптации: Концепция применима для любых взаимодействий с LLM, где требуется не только информация, но и принятие решений или выполнение действий.

Архитектура агентных систем:

Прямая применимость: Ограниченная для обычных пользователей, но понимание компонентов агента помогает эффективнее взаимодействовать с системой. Концептуальная ценность: Средняя. Понимание работы памяти и планирования помогает пользователям строить более эффективные запросы. Потенциал для адаптации: Знание о компонентах агента позволяет пользователям "подсказывать" системе использовать память о предыдущих взаимодействиях или планировать действия.

Практические применения:

Прямая применимость: Высокая. Примеры персонального и бизнес-ассистентов наглядно демонстрируют, как можно применять агентный подход в повседневных задачах. Концептуальная ценность: Высокая. Примеры помогают пользователям представить, как можно применять агентный подход для решения своих задач. Потенциал для адаптации: Отличный. Пользователи могут адаптировать показанные сценарии для своих нужд.

Оценка и оптимизация систем:

Прямая применимость: Низкая для обычных пользователей, но полезно для понимания ограничений систем. Концептуальная ценность: Средняя. Понимание метрик помогает пользователям формировать реалистичные ожидания. Потенциал для адаптации: Ограниченный, в основном полезен для разработчиков.

Prompt:

Использование концепции агентного информационного поиска в промтах для GPT Исследование агентного информационного поиска (Agentic IR) предлагает новый подход к взаимодействию с языковыми моделями, переходя от простого получения

информации к динамическому достижению информационных состояний через серию целенаправленных действий.

Ключевые концепции для применения в промтах

Динамические информационные состояния вместо статических запросов **Модульность агента** (профиль, память, планирование, действие) **Итеративное уточнение запросов** и адаптация к обратной связи **Проактивное планирование** для достижения информационных целей ## Пример промта, использующего принципы агентного IR

[======] # Задача: Помощь в планировании деловой поездки в Сингапур

Профиль агента Ты - бизнес-ассистент с возможностями агентного информационного поиска. Твоя цель - не просто предоставить информацию, а помочь мне достичь целевого информационного состояния для успешной деловой поездки.

##