Многоисточниковая обрезка знаний для генерации с учетом извлечения: Бенчмарк и эмпирическое исследование

Дата: 2025-02-16 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2409.13694

Рейтинг: 62
Адаптивность: 75

Ключевые выводы:

Исследование направлено на разработку и оценку фреймворка PruningRAG для улучшения работы с несколькими источниками знаний в системах Retrieval Augmented Generation (RAG). Основные результаты показывают, что многоуровневая стратегия отсечения нерелевантной информации из разнородных источников значительно повышает точность ответов и снижает уровень галлюцинаций в LLM.

Объяснение метода:

Исследование представляет ценную концепцию фильтрации разнородных источников знаний и методы улучшения рассуждений LLM (CoT, ICL). Пользователи могут применить принципы выбора надежных источников и пошагового рассуждения, но полная реализация технически сложна. Основная ценность - в понимании работы с противоречивой информацией и структурирования запросов для получения более точных ответов.

Ключевые аспекты исследования: 1. **Multi-Source Knowledge Pruning (Pruning RAG)** - исследование представляет новый фреймворк, который использует многоуровневую фильтрацию знаний из разных источников для улучшения работы систем RAG (Retrieval-Augmented Generation).

Двухуровневая фильтрация - метод включает в себя крупнозернистую фильтрацию (отбор релевантных источников знаний) и мелкозернистую фильтрацию (уточнение контекста внутри выбранных источников).

Структурированный бенчмарк - авторы стандартизировали набор данных, содержащий разнородные источники знаний (веб-страницы и API), для оценки эффективности RAG-систем в реальных сценариях.

CoT и ICL для рассуждений - исследование показывает, как использование

Chain-of-Thought (цепочка рассуждений) и In-Context Learning (обучение в контексте) улучшает качество ответов при работе с отфильтрованными знаниями.

Эмпирический анализ гиперпараметров - авторы провели детальное исследование влияния размера чанков, их перекрытия и количества на производительность системы.

Дополнение: Для работы методов из исследования действительно требуется определенная техническая инфраструктура, включая доступ к API и возможность дообучения моделей. Однако многие концептуальные идеи могут быть адаптированы для использования в стандартном чате без технических модификаций:

Двухуровневая стратегия проверки информации - пользователь может сначала спросить LLM о надежных источниках для ответа на вопрос, а затем уточнить информацию из этих источников. Например: "Какие источники лучше всего подходят для информации о [тема]?" и затем "Предоставь информацию о [конкретный вопрос] из [названные источники]".

Chain-of-Thought (CoT) - можно применить без модификаций, просто попросив модель "рассуждать шаг за шагом" или "объяснить процесс рассуждения".

In-Context Learning с примерами из других доменов - пользователь может предоставить примеры из другой области, чтобы показать желаемый формат ответа. Исследование показало, что примеры из разных доменов снижают эффект переобучения и улучшают способность модели критически оценивать информацию.

Баланс объема контекста - исследование показывает, что умеренный объем контекста (не слишком мало и не слишком много) дает лучшие результаты. Пользователи могут фокусировать свои запросы, избегая информационной перегрузки.

Работа с противоречивой информацией - пользователь может явно попросить модель сравнить информацию из разных источников и объяснить противоречия.

Результаты от применения этих концепций: - Снижение галлюцинаций и повышение точности ответов - Улучшение способности модели критически оценивать информацию - Более структурированные и понятные ответы - Повышение уверенности пользователя в полученной информации

Важно отметить, что хотя техническая реализация полного фреймворка PruningRAG требует специальных навыков, основные принципы работы с множественными источниками информации могут быть применены любым пользователем в стандартном чате с LLM.

Анализ практической применимости: 1. **Multi-Source Knowledge Pruning** - Прямая применимость: Средняя. Концепция фильтрации источников может быть применена пользователями при формулировании запросов к LLM, но полная

реализация требует технических знаний. - Концептуальная ценность: Высокая. Понимание того, что объединение разных источников знаний может привести к противоречиям и галлюцинациям, помогает пользователям более критично относиться к ответам LLM. - Потенциал для адаптации: Высокий. Идея проверки информации из нескольких источников и отсеивания противоречивых данных может быть применена пользователями даже без технической реализации.

Двухуровневая фильтрация Прямая применимость: Низкая. Технически сложно реализовать для обычного пользователя. Концептуальная ценность: Высокая. Понимание необходимости сначала выбрать правильные источники, а затем уточнить информацию внутри них, помогает пользователям формулировать более эффективные запросы. Потенциал для адаптации: Средний. Пользователи могут адаптировать этот принцип, явно указывая в запросе, какие источники предпочтительнее и какую информацию они хотят получить.

CoT и ICL для рассуждений

Прямая применимость: Высокая. Пользователи могут непосредственно применять принципы пошагового рассуждения и обучения на примерах в своих запросах. Концептуальная ценность: Высокая. Понимание того, как формулировать запросы для получения пошаговых рассуждений и использовать примеры из других доменов, значительно улучшает взаимодействие с LLM. Потенциал для адаптации: Высокий. Эти методы могут быть легко адаптированы пользователями в повседневных запросах.

Анализ гиперпараметров

Прямая применимость: Низкая. Обычные пользователи не могут напрямую контролировать размер чанков или их перекрытие. Концептуальная ценность: Средняя. Понимание того, что избыток контекста может ухудшить ответ, помогает пользователям быть более лаконичными и точными в запросах. Потенциал для адаптации: Средний. Пользователи могут адаптировать принцип "умеренного объема контекста" при формулировании запросов.

Структурированный бенчмарк

Прямая применимость: Низкая. Набор данных предназначен для исследователей, а не конечных пользователей. Концептуальная ценность: Средняя. Понимание разнообразия источников знаний помогает пользователям осознать ограничения LLM. Потенциал для адаптации: Низкий. Относится в основном к исследовательской деятельности.

Prompt:

Использование знаний из исследования PruningRAG в промптах для GPT ## Ключевое понимание исследования

Исследование PruningRAG показывает, что двухуровневая стратегия отсечения

нерелевантной информации из разных источников значительно улучшает точность ответов и снижает галлюцинации в LLM.

Пример промпта с применением знаний из исследования

[======] # Промпт для анализа информации из нескольких источников

Я предоставлю тебе информацию из нескольких источников (веб-страницы и структурированные данные API) по теме [ТЕМА]. Пожалуйста, используй двухуровневый подход обработки:

Этап 1: Грубое отсечение источников - Оцени релевантность каждого источника к моему вопросу - Отбрось полностью нерелевантные источники - Обозначь, какие источники ты сохранил и почему

Этап 2: Тонкое отсечение содержимого - В выбранных источниках выдели только релевантные фрагменты - Для веб-страниц используй комбинацию семантического поиска и ключевых слов - Для структурированных данных сфокусируйся на конкретных сущностях

Этап 3: Формирование ответа - Используй рассуждение по цепочке мыслей (CoT) для неструктурированных источников - Будь более прямолинейным при работе со структурированными данными - Синтезируй ответ размером 200-500 токенов для оптимального баланса полноты и фокуса

Вопрос: [ВОПРОС]

Источники: 1. [ИСТОЧНИК 1 - веб-страница] 2. [ИСТОЧНИК 2 - API данные] 3. [ИСТОЧНИК 3 - веб-страница] ... [======]

Как работают знания из исследования в этом промпте

Двухуровневая стратегия отсечения - промпт явно запрашивает сначала отбросить нерелевантные источники целиком, а затем отфильтровать содержимое в оставшихся источниках, что соответствует ключевой методологии PruningRAG.

Дифференцированный подход к типам источников - промпт учитывает разницу между структурированными (API) и неструктурированными (веб) источниками, что отражает адаптивную методологию исследования.

Оптимальный размер ответа - запрос на синтез информации в объеме 200-500 токенов соответствует выводам исследования об оптимальном размере чанков.

Избирательное применение CoT - промпт предлагает использовать рассуждение по цепочке мыслей для неструктурированных данных, но быть более прямолинейным для структурированных источников, что согласуется с результатами исследования.

Такой подход к формированию промпта позволяет значительно повысить точность ответов GPT и снизить вероятность галлюцинаций при работе с множественными источниками данных.