Множественный уровень абстракции для извлечения и увеличения генерации

Дата: 2025-01-28 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2501.16952

Рейтинг: 72
Адаптивность: 85

Ключевые выводы:

Исследование предлагает новый подход к Retrieval Augmented Generation (RAG) под названием Multiple Abstraction Level RAG (MAL-RAG), который использует иерархическую структуру документов для извлечения информации на разных уровнях абстракции (документ, раздел, параграф, предложение). Основная цель - улучшить точность ответов на вопросы в научных доменах, особенно в области гликонауки. Результаты показывают, что MAL-RAG превосходит традиционные подходы RAG на 25,739% по метрике корректности ответов.

Объяснение метода:

Исследование предлагает ценную концепцию многоуровневой абстракции для RAG-систем, которая помогает решить проблему "lost in the middle" и улучшает качество ответов. Хотя полная реализация требует технических знаний, основные принципы могут быть адаптированы обычными пользователями для структурирования запросов на разных уровнях детализации.

Ключевые аспекты исследования: 1. **Multiple Abstraction Level (MAL) подход** - исследование представляет новую технику RAG (Retrieval Augmented Generation), которая использует иерархическую многоуровневую структуру документов для извлечения информации на разных уровнях абстракции: уровень всего документа, уровень раздела, уровень абзаца и уровень нескольких предложений.

Решение проблемы "lost in the middle" - авторы предлагают способ преодоления проблемы, когда LLM теряет внимание к информации в середине длинного контекста, путем использования более компактных высокоуровневых абстракций.

Map-reduce подход к суммаризации - для создания индексов документов и разделов применяется многоэтапное суммирование: сначала суммируются абзацы, затем из этих саммари создаются суммаризации разделов, и наконец - суммаризации целых документов.

Вероятностный отбор чанков - система использует пороговый механизм для

отбора наиболее релевантных чанков, преобразуя оценки сходства в вероятности с помощью softmax и отбирая чанки до достижения заданного порога кумулятивной вероятности.

Экспериментальная валидация на научных текстах - метод был протестирован на специализированной научной области (гликонауке) и показал существенное улучшение качества ответов по сравнению с традиционными RAG-подходами, использующими чанки одного уровня абстракции.

Дополнение: Действительно ли для работы методов этого исследование требуется дообучение или API? Или методы и подходы можно применить в стандартном чате, а ученые лишь для своего удобства использовали расширенные техники?

Для полной реализации MAL-RAG в том виде, как описано в статье, требуется некоторая техническая инфраструктура, включая: 1. Система индексации документов на разных уровнях абстракции 2. Модель для генерации суммаризаций (авторы использовали Vicuna-13B) 3. Система поиска по индексам с расчетом косинусного сходства

Однако ключевые концепции и подходы можно адаптировать для использования в стандартном чате без дополнительных API или дообучения:

Многоуровневые запросы: Пользователь может последовательно запрашивать информацию на разных уровнях абстракции: "Дай краткий обзор темы X" (уровень документа) "Расскажи подробнее о разделе Y в теме X" (уровень раздела) "Объясни детально процесс Z из раздела Y" (уровень абзаца/предложений)

Преодоление "lost in the middle": Пользователь может разбивать сложные запросы на более короткие, связанные части, чтобы LLM мог сфокусироваться на каждой части отдельно:

Сначала запрос общей информации Затем серия конкретных вопросов по деталям

Структурированная суммаризация: Пользователь может запрашивать поэтапную суммаризацию информации:

"Суммируй ключевые моменты из X" "Теперь объедини эти ключевые моменты в общую картину"

Осознанное использование контекста: Понимая, что LLM имеет ограничения в обработке длинного контекста, пользователь может:

Явно указывать, какая информация наиболее важна Просить модель сначала обработать информацию, а затем ответить на вопрос Результаты применения этих концепций в стандартном чате: - Более структурированные и точные ответы - Лучшее понимание сложных тем благодаря представлению информации на разных уровнях детализации - Снижение проблемы "lost in the middle" за счет фокусировки

внимания модели - Более эффективное использование контекстного окна модели

Таким образом, хотя полная техническая реализация MAL-RAG требует дополнительной инфраструктуры, основные концептуальные принципы могут быть успешно применены в стандартном чате с LLM.

Анализ практической применимости: **Multiple Abstraction Level подход**: - Прямая применимость: Средняя. Обычные пользователи не могут напрямую реализовать полноценную MAL-RAG систему без технических знаний, но концепция использования иерархической информации может быть адаптирована для структурирования запросов. - Концептуальная ценность: Высокая. Понимание того, что информация имеет разные уровни абстракции, помогает пользователям формулировать более эффективные запросы, запрашивая как общую информацию, так и конкретные детали. - Потенциал для адаптации: Высокий. Пользователи могут вручную запрашивать информацию на разных уровнях детализации, имитируя многоуровневый подход.

Решение проблемы "lost in the middle": - Прямая применимость: Средняя. Пользователи могут применять технику разбиения длинных запросов на более короткие, структурированные части. - Концептуальная ценность: Высокая. Понимание ограничений LLM в обработке длинного контекста помогает пользователям оптимизировать свои запросы. - Потенциал для адаптации: Высокий. Пользователи могут создавать многоэтапные запросы, начиная с общих вопросов и постепенно углубляясь в детали.

Map-reduce подход к суммаризации: - Прямая применимость: Низкая. Сложен для реализации обычными пользователями без программирования. - Концептуальная ценность: Средняя. Понимание принципа постепенного обобщения информации может помочь в структурировании запросов. - Потенциал для адаптации: Средний. Пользователи могут запрашивать последовательную суммаризацию информации, начиная с деталей и переходя к более общим выводам.

Вероятностный отбор чанков: - Прямая применимость: Низкая. Требует технической реализации. - Концептуальная ценность: Средняя. Понимание важности фильтрации информации по релевантности. - Потенциал для адаптации: Средний. Пользователи могут явно указывать приоритеты в своих запросах.

Экспериментальная валидация на научных текстах: - Прямая применимость: Средняя. Результаты показывают эффективность подхода для работы со сложными научными текстами. - Концептуальная ценность: Высокая. Демонстрирует, что структурированный подход к извлечению информации значительно улучшает качество ответов. - Потенциал для адаптации: Высокий. Подход может быть адаптирован для работы с любыми сложными текстами, не только научными.

Prompt:

Использование MAL-RAG в промптах для GPT ## Как применять знания из исследования

Исследование MAL-RAG (Multiple Abstraction Level RAG) предлагает иерархический подход к извлечению информации на разных уровнях абстракции. Это знание можно эффективно применить при составлении промптов для GPT.

Пример промпта с использованием MAL-RAG концепции

[======] Проанализируй следующую научную статью по биохимии, используя многоуровневый подход к извлечению информации:

[ВСТАВИТЬ ТЕКСТ СТАТЬИ]

Сначала дай общее резюме всего документа (уровень документа). Затем выдели ключевые разделы и их основные идеи (уровень раздела). Для наиболее важных разделов предоставь детальный анализ ключевых параграфов (уровень параграфа). Наконец, выдели 5-7 критически важных предложений, содержащих основные выводы или методологические инновации (уровень предложения). В своем ответе используй map-reduce подход: для каждого уровня создавай сжатую версию, которая сохраняет ключевую информацию, но устраняет избыточность. Обрати особое внимание на методологические детали и количественные результаты. [======]

Почему это работает

Этот промпт работает эффективно, потому что:

Использует иерархическую структуру - следуя принципам MAL-RAG, промпт запрашивает анализ на четырех уровнях абстракции **Применяет map-reduce подход** - просит создавать сжатые версии на каждом уровне, что помогает избежать проблемы "lost in the middle" **Фокусируется на естественной структуре** - использует логическую структуру документа вместо произвольного разделения **Обеспечивает полноту охвата** - гарантирует, что будет захвачена информация как общего характера, так и конкретные детали Такой подход позволяет получить более точные, полные и структурированные ответы от GPT, особенно при работе со сложными научными текстами, где важно не упустить ключевые детали, но при этом сохранить общий контекст.