Влияние размера контекста и выбора модели в системах генерации с дополнением информации из поиска

Дата: 2025-02-20 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2502.14759

Рейтинг: 65
Адаптивность: 75

Ключевые выводы:

Исследование направлено на систематическое изучение влияния размера контекста, выбора базовой языковой модели и метода поиска на эффективность систем генерации с дополнением поиском (RAG). Основные результаты показывают, что производительность RAG-систем улучшается с увеличением количества контекстных фрагментов до 10-15, после чего наблюдается стагнация или снижение эффективности. Также выявлено, что разные модели показывают лучшие результаты в разных доменах: Mistral и Qwen лучше работают с биомедицинскими данными, а GPT и Llama - с энциклопедическими.

Объяснение метода:

Исследование предоставляет ценные рекомендации по оптимальному количеству контекста (10-15 фрагментов) и выбору моделей для разных доменов. Пользователи могут адаптировать эти принципы для структурирования запросов и выбора моделей. Однако многие аспекты требуют технической экспертизы и прямого доступа к компонентам RAG-систем, что ограничивает применимость для обычных пользователей.

Ключевые аспекты исследования: 1. **Влияние размера контекста на эффективность RAG-систем**: Исследование систематически анализирует, как количество контекстных фрагментов (от 1 до 30) влияет на качество ответов на вопросы. Результаты показывают, что производительность улучшается до примерно 15 фрагментов, после чего наступает стагнация или даже снижение.

Сравнение различных базовых LLM в RAG-системах: Авторы тестируют 8 различных моделей (GPT-3.5, GPT-4o, LLaMa3, Mixtral, Qwen и другие) на двух разных доменах - биомедицинском (BioASQ) и энциклопедическом (QuoteSum). Результаты показывают, что Mixtral и Qwen лучше работают с биомедицинскими данными, а GPT и LLaMa - с энциклопедическими.

Сравнение методов извлечения информации: Исследование сравнивает два метода поиска - семантический поиск и BM25 (поиск на основе ключевых слов). BM25 показывает лучшие результаты для биомедицинских данных, оптимизируя точность поиска.

Конфликт между внутренним знанием LLM и внешним контекстом: Авторы обнаружили, что в некоторых случаях внутренние знания модели могут дать более точный ответ, чем ответ, основанный на извлеченных фрагментах контекста, особенно если поиск возвращает нерелевантные результаты.

Открытый и закрытый поиск информации: Исследование сравнивает эффективность RAG-систем в сценариях с заранее известными релевантными фрагментами и в реалистичных условиях, когда система должна самостоятельно искать информацию в больших корпусах документов.

Дополнение:

Применимость методов исследования в стандартном чате

Исследование не требует дообучения или специального API для применения основных концепций. Хотя авторы использовали различные модели и методы поиска для своих экспериментов, ключевые выводы можно адаптировать для стандартного чата.

Концепции, применимые в стандартном чате:

Оптимальное количество контекста: Пользователь может ограничивать объем предоставляемой информации до 10-15 ключевых фрагментов, избегая информационной перегрузки модели.

Структурирование контекста: Размещение наиболее важной информации в начале и конце запроса, учитывая эффект "потери в середине".

Специфичность запросов: Формулирование запросов с конкретными ключевыми словами вместо общих семантических запросов для повышения точности ответов.

Сравнение ответов с контекстом и без: Пользователь может задать один и тот же вопрос с предоставлением контекста и без него, чтобы сравнить, как модель использует внутренние знания и внешнюю информацию.

Выбор модели под задачу: При наличии доступа к разным моделям, пользователь может выбирать специфические модели для разных доменов (например, Mixtral для биомедицинских вопросов).

Ожидаемые результаты от применения:

• Более точные и релевантные ответы

• Снижение вероятности "галлюцинаций" из-за информационного шума

• Лучшее использование внутренних знаний модели при необходимости

• Повышение эффективности взаимодействия с LLM за счет оптимальной структуры запросов

Анализ практической применимости: **Влияние размера контекста** - Прямая применимость: Пользователи могут оптимизировать количество контекстных фрагментов в своих RAG-системах, ориентируясь на 10-15 фрагментов как оптимальное количество. Это конкретный, практически применимый вывод для любого, кто настраивает RAG-системы. - Концептуальная ценность: Понимание, что "больше не всегда лучше" в контексте RAG-систем помогает пользователям осознать, как LLM обрабатывают информацию и почему слишком много контекста может быть вредным. - Потенциал для адаптации: Пользователи могут использовать принцип "оптимального размера контекста" даже в обычных чатах, структурируя свои запросы с ограниченным, но достаточным количеством информации.

Сравнение базовых LLM - Прямая применимость: Пользователи могут выбирать конкретные модели в зависимости от домена их задач (например, Mixtral для биомедицинских задач). - Концептуальная ценность: Исследование демонстрирует, что размер модели не всегда определяет ее эффективность — Mixtral (8x7B) и Qwen (7B) превосходят более крупные модели в биомедицинских задачах. - Потенциал для адаптации: Пользователи могут применять эти знания при выборе моделей для своих задач, понимая их сильные и слабые стороны.

Методы извлечения информации - Прямая применимость: Выбор BM25 для биомедицинских задач может улучшить точность ответов. - Концептуальная ценность: Понимание различий между семантическим поиском (широкий охват) и поиском на основе ключевых слов (точность) помогает пользователям лучше формулировать запросы. - Потенциал для адаптации: Пользователи могут адаптировать свои запросы, делая их более конкретными и ориентированными на ключевые слова для получения более точных ответов.

Конфликт знаний - Прямая применимость: Ограниченная для обычного пользователя, так как сложно контролировать, когда модель использует внутренние знания, а когда — предоставленный контекст. - Концептуальная ценность: Высокая — понимание, что иногда внутренние знания модели могут быть лучше, чем неточно найденный контекст, помогает критически оценивать ответы. - Потенциал для адаптации: Пользователи могут экспериментировать с запросами с контекстом и без него, сравнивая ответы для получения наиболее полной информации.

Открытый и закрытый поиск - Прямая применимость: Понимание ограничений открытого поиска помогает пользователям формировать реалистичные ожидания от RAG-систем. - Концептуальная ценность: Осознание сложности поиска релевантной

информации в больших корпусах документов. - Потенциал для адаптации: Пользователи могут улучшать свои запросы, делая их более специфичными и точными для повышения релевантности поиска.

Prompt:

Использование знаний из исследования о RAG-системах в промптах для GPT ## Ключевые выводы для применения в промптах

Исследование о влиянии размера контекста и выборе моделей в RAG-системах предоставляет ценные знания, которые можно применить при составлении эффективных промптов:

Оптимальное количество контекстных фрагментов: 10-15 фрагментов **Специализация моделей по доменам**: разные модели лучше работают с разными типами данных **Важность точного поиска**: BM25 часто превосходит семантический поиск **Баланс между внешними данными и внутренними знаниями модели** ## Пример промпта с применением знаний из исследования

[======] # Запрос для медицинской информации с оптимизированным контекстом

Контекст (ограничен 12 релевантными фрагментами) [Здесь размещаются 10-12 релевантных фрагментов из надежных медицинских источников]

Инструкции Ты работаешь как медицинский исследовательский ассистент. Используя предоставленные контекстные фрагменты:

Ответь на следующий вопрос о лечении диабета 2 типа новыми препаратами. Если в контексте недостаточно информации, укажи это явно, но предложи ответ на основе своих базовых знаний. Четко разграничь информацию из предоставленного контекста и свои базовые знания. Обрати особое внимание на фрагменты 3, 5 и 8, которые содержат ключевую информацию по теме. ## Вопрос Какие новые GLP-1 агонисты показывают наилучшие результаты в снижении сердечно-сосудистых рисков у пациентов с диабетом 2 типа? [======]

Объяснение эффективности этого подхода

Оптимальное количество контекста: В промпте используется 10-12 фрагментов, что соответствует оптимальному диапазону (10-15), выявленному в исследовании.

Специализация по домену: Промпт явно указывает на биомедицинскую тематику, где модели типа Mistral и Qwen показывают лучшие результаты.

Приоритизация контекста: В промпте указаны наиболее важные фрагменты (3, 5, 8), что помогает модели сфокусироваться на ключевой информации и избежать "потери информации в середине".

Гибридный подход: Промпт предлагает модели использовать как предоставленный контекст, так и внутренние знания, когда это необходимо, что соответствует выводам исследования о целесообразности комбинированного подхода.

Явное разграничение источников: Требование разделять информацию из контекста и базовые знания модели помогает контролировать качество и происхождение информации.

Такой подход к составлению промптов позволяет максимально эффективно использовать возможности GPT, учитывая научно обоснованные ограничения и особенности работы RAG-систем.