Ворота контекстной осведомленности для увеличенной генерации извлечения

Дата: 2025-01-06 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2411.16133

Рейтинг: 68
Адаптивность: 75

Ключевые выводы:

Исследование направлено на решение проблемы извлечения нерелевантной информации в системах Retrieval Augmented Generation (RAG). Авторы предлагают новую архитектуру Context Awareness Gate (CAG), которая динамически корректирует входной промпт для LLM в зависимости от того, требует ли запрос пользователя извлечения внешнего контекста. Результаты показывают, что CAG значительно улучшает релевантность контекста и качество ответов в системах открытого доменного вопросно-ответного взаимодействия.

Объяснение метода:

Исследование предлагает метод динамического определения необходимости внешнего контекста для запросов к LLM, что повышает точность ответов. Концепция адаптируема для обычных пользователей в виде инструкций в промпте, хотя полная реализация требует технических навыков. Работа решает фундаментальную проблему взаимодействия с LLM, предлагая статистически обоснованный подход.

Ключевые аспекты исследования: 1. **Context Awareness Gate (CAG)** - новая архитектура, которая динамически корректирует входной промпт LLM на основе анализа необходимости использования внешнего контекста для ответа на вопрос пользователя. Система определяет, нужно ли извлекать информацию из внешней базы данных или LLM может ответить, используя свои внутренние знания.

Vector Candidates (VC) - статистический метод, который анализирует распределение эмбеддингов контекста и псевдо-запросов для определения релевантности запроса пользователя к имеющейся базе знаний. Метод не требует использования LLM для классификации запросов, что делает его более эффективным и масштабируемым.

Статистический анализ распределений - авторы исследовали распределения косинусной схожести между контекстами и запросами, определив чёткие статистические различия между релевантными и нерелевантными парами, что позволяет эффективно классифицировать запросы.

Context Retrieval Supervision Benchmark (CRSB) - новый набор данных из 17 различных тематик, предназначенный для оценки эффективности систем, основанных на контекстной осведомленности и семантической маршрутизации.

Дополнение: Для работы методов этого исследования в полном объеме действительно требуется доступ к API для работы с эмбеддингами и некоторая техническая настройка. Однако ключевые концепции и подходы вполне можно адаптировать для использования в стандартном чате, без необходимости дообучения моделей.

Концепции, которые можно применить в стандартном чате:

Динамическое переключение между режимами ответа - пользователь может включать в свои промпты инструкции вида: "Сначала определи, требует ли мой вопрос специализированных знаний, которыми ты, возможно, не обладаешь. Если да, сообщи мне об этом. Если нет, ответь, используя свои знания."

Предварительная классификация запросов - пользователь может сам определять тип своего запроса: "Это общий вопрос, на который ты должен знать ответ" или "Это специфический вопрос, для которого может потребоваться дополнительная информация".

Мета-запросы для определения уверенности - перед основным вопросом пользователь может спросить: "Насколько ты уверен в своих знаниях о [тема]?", что поможет определить необходимость внешней информации.

Цепочка рассуждений для самопроверки - можно попросить модель использовать подход "цепочки рассуждений", чтобы она сама определила, достаточно ли у неё знаний: "Рассуждай шаг за шагом, чтобы определить, достаточно ли у тебя информации для ответа на этот вопрос".

Результаты от применения этих концепций: 1. Более точные ответы, так как модель будет ясно сообщать о пробелах в своих знаниях 2. Снижение галлюцинаций в ответах на вопросы, выходящие за рамки знаний модели 3. Более эффективное взаимодействие, поскольку пользователь будет понимать, когда ему нужно предоставить дополнительный контекст 4. Повышение доверия к ответам модели благодаря явному разделению между уверенными и неуверенными ответами

Эти адаптации не требуют технических навыков и могут быть реализованы в обычном чате любым пользователем.

Анализ практической применимости: 1. **Context Awareness Gate (CAG)** - **Прямая применимость**: Высокая. Обычные пользователи могут адаптировать принцип CAG для своих запросов, разделяя их на те, где нужна внешняя информация, и те, где достаточно знаний модели. Это позволит получать более точные ответы и экономить ресурсы. - **Концептуальная ценность**: Очень высокая. Понимание

разницы между запросами, требующими внешних данных, и теми, на которые LLM может ответить самостоятельно, помогает пользователям формулировать более эффективные запросы. - **Потенциал для адаптации**: Высокий. Пользователи могут внедрить логику принятия решений в свои промпты, например: "Если ты знаешь ответ на этот вопрос, ответь напрямую. Если нет, скажи, что тебе нужна дополнительная информация."

Vector Candidates (VC) Прямая применимость: Средняя. Хотя сам алгоритм требует технических знаний, концепция проверки релевантности запроса к имеющейся базе знаний может быть использована даже неспециалистами. **Концептуальная ценность**: Высокая. Метод демонстрирует, как можно эффективно определять, когда запрос выходит за рамки конкретной области знаний. **Потенциал для адаптации**: Средний. Пользователи могут адаптировать принцип, предварительно формулируя свои запросы так, чтобы указать, требуется ли специфическая информация или достаточно общих знаний.

Статистический анализ распределений

Прямая применимость: Низкая для обычных пользователей, высокая для разработчиков. **Концептуальная ценность**: Средняя. Понимание того, как определяется релевантность запроса к контексту, помогает лучше формулировать запросы. **Потенциал для адаптации**: Средний. Пользователи могут использовать знания о пороговых значениях релевантности для оптимизации своих запросов.

CRSB Dataset

Прямая применимость: Низкая для обычных пользователей. **Концептуальная ценность**: Средняя. Демонстрирует разнообразие тематик и типов запросов. **Потенциал для адаптации**: Низкий. В основном представляет ценность для исследователей и разработчиков.

Prompt:

Применение исследования CAG в промптах для GPT ## Ключевые знания из отчета, полезные для промптов

• Context Awareness Gate (CAG) - архитектура, определяющая необходимость извлечения внешнего контекста

• Vector Candidates (VC) - статистический метод анализа эмбеддингов для определения релевантности контекста

• Пороговые значения сходства: >0.55 для релевантных пар, <0.21 для нерелевантных

Пример промпта с применением знаний из исследования

[======] # Запрос с контекстной осведомленностью

Контекст [Вставьте здесь ваши документы или базу знаний]

Инструкции Ты - ассистент с улучшенной контекстной осведомленностью. Используя принципы Context Awareness Gate:

Проанализируй мой вопрос и определи, требует ли он внешних знаний из предоставленного контекста. Если косинусное сходство между моим вопросом и контекстом оценивается выше 0.55, используй информацию из контекста. Если сходство ниже 0.21, полагайся на свои внутренние знания. В пограничных случаях (0.21-0.55) явно укажи источник своих знаний и уровень уверенности. ## Вопрос [Вставьте здесь ваш вопрос] [======]

Как это работает

Динамическая оценка необходимости контекста: Промпт инструктирует GPT симулировать работу CAG, оценивая релевантность контекста к запросу.

Использование пороговых значений: Применяются научно обоснованные пороговые значения из исследования (0.55 и 0.21).

Прозрачность источников: GPT указывает, опирается ли он на предоставленный контекст или на собственные знания.

Оптимизация ресурсов: Контекст используется только когда он действительно релевантен, что улучшает качество ответов и экономит токены.

Этот подход позволяет создать более "умную" RAG-систему, которая не просто извлекает информацию из контекста для каждого запроса, а делает это избирательно, повышая релевантность ответов.