Измерение и повышение доверия к LLM в RAG через обоснованные атрибуции и обучение отказу

Дата: 2025-03-03 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2409.11242

Рейтинг: 75
Адаптивность: 70

Ключевые выводы:

Исследование направлено на измерение и улучшение надежности больших языковых моделей (LLM) в системах генерации с дополнением из поиска (RAG) через обоснованные атрибуции. Авторы представили метрику TRUST-SCORE для оценки надежности LLM и метод TRUST-ALIGN для улучшения этой надежности, который значительно превзошел базовые методы на нескольких наборах данных.

Объяснение метода:

Исследование вводит важные метрики и методы для повышения надежности LLM в RAG-системах. Концепции TRUST-SCORE и понимание типов галлюцинаций имеют высокую практическую ценность для пользователей. Хотя полная реализация TRUST-ALIGN требует технических навыков, принципы могут быть адаптированы для улучшения взаимодействия с LLM и критической оценки их ответов.

Ключевые аспекты исследования: 1. **Метрика TRUST-SCORE** - комплексный показатель для оценки надежности и достоверности LLM в контексте RAG-систем, который оценивает: а) способность модели отказаться от ответа при недостатке информации, б) точность ответов на основе документов, в) обоснованность цитирования источников.

Метод TRUST-ALIGN - подход для улучшения надежности LLM в RAG путем создания специального набора данных (19 тыс. примеров) и обучения моделей с помощью Direct Preference Optimization (DPO). Метод фокусируется на исправлении пяти типов галлюцинаций в RAG-системах.

Выявление проблемы параметрического знания - исследование показывает, что современные LLM (включая GPT-4) чрезмерно полагаются на внутренние параметрические знания вместо предоставленных документов, что снижает их эффективность в RAG-системах.

Улучшение способности отказа от ответа - значительное повышение способности моделей корректно отказываться от ответа, когда в предоставленных документах

недостаточно информации для ответа на вопрос.

Повышение качества цитирования - улучшение способности моделей обосновывать свои утверждения ссылками на релевантные документы.

Дополнение:

Да, для работы методов этого исследования в полной мере требуется дообучение моделей и использование специализированного API. Однако многие концепции и подходы можно адаптировать для применения в стандартном чате.

Применимые концепции и подходы:

Структура запросов с требованием цитирования Явно просить модель подтверждать свои утверждения ссылками на конкретные части предоставленного контекста Пример: "Ответь на вопрос, используя только предоставленную информацию, и укажи, из какого абзаца ты взял каждый факт"

Проверка обоснованности цитирования

Самостоятельно проверять, соответствуют ли утверждения модели указанным источникам Использовать указанную в исследовании концепцию F1_GC (точность и полнота цитирования)

Запрос на отказ от ответа

Явно инструктировать модель отказываться от ответа, если в предоставленных документах недостаточно информации Пример: "Если в предоставленных документах недостаточно информации для ответа, пожалуйста, напиши: 'Недостаточно информации для ответа на этот вопрос'"

Разделение параметрического и документального знания

Просить модель четко разграничивать информацию из предоставленных документов и общие знания Пример: "Укажи, какая информация взята из предоставленных документов, а какая основана на общих знаниях"

Применение компонентов TRUST-SCORE для самооценки

Просить модель оценить свою уверенность в ответе Запрашивать обоснование каждого сделанного утверждения ### Ожидаемые результаты:

• Повышение прозрачности ответов модели

• Снижение риска необоснованных утверждений

• Более критичный подход к оценке ответов LLM

• Повышение доверия к обоснованным ответам

• Лучшее понимание границ знаний модели

Хотя эти адаптации не дадут таких значительных улучшений, как полное дообучение по методу TRUST-ALIGN, они могут существенно повысить качество взаимодействия с LLM в стандартном чате и помочь пользователям лучше оценивать надежность получаемой информации.

Анализ практической применимости: **1. Метрика TRUST-SCORE** - **Прямая применимость**: Высокая. Пользователи, особенно разработчики и оценщики RAG-систем, могут применять эту метрику для оценки качества и надежности ответов, получаемых от LLM. Подход к оценке может быть адаптирован даже для ручной проверки ответов. - **Концептуальная ценность**: Очень высокая. Метрика помогает понять многомерную природу надежности LLM, разделяя её на компоненты: способность отказываться от ответа, точность ответов и качество цитирования. - **Потенциал для адаптации**: Высокий. Компоненты метрики могут быть использованы отдельно для оценки конкретных аспектов работы LLM, применимы к различным моделям и сценариям использования.

2. Метод TRUST-ALIGN - **Прямая применимость**: Средняя. Рядовые пользователи не смогут самостоятельно реализовать этот метод, так как он требует создания специального набора данных и обучения моделей. Однако разработчики могут внедрить этот подход для улучшения своих RAG-систем. - **Концептуальная ценность**: Высокая. Демонстрирует, как можно существенно улучшить модели для работы в RAG-системах, фокусируясь на конкретных типах галлюцинаций. - **Потенциал для адаптации**: Средний. Принципы создания обучающих данных могут быть адаптированы для других задач, где требуется улучшение надежности LLM.

3. Выявление проблемы параметрического знания - **Прямая применимость**: Высокая. Пользователи должны понимать, что ответы LLM могут основываться на внутренних знаниях, а не на предоставленных документах, что важно для критической оценки получаемых ответов. - **Концептуальная ценность**: Высокая. Осознание этой проблемы помогает пользователям формулировать более эффективные запросы и критически оценивать ответы моделей. - **Потенциал для адаптации**: Средний. Понимание этой проблемы может быть использовано для разработки лучших практик взаимодействия с LLM.

4. Улучшение способности отказа от ответа - **Прямая применимость**: Высокая. Пользователи могут применять специальные промпты, чтобы побудить модель отказываться от ответа при недостатке информации, хотя исследование показывает ограниченную эффективность такого подхода без специального обучения. - **Концептуальная ценность**: Высокая. Понимание важности способности модели признавать незнание критично для надежного использования LLM. - **Потенциал для адаптации**: Высокий. Концепция может быть адаптирована для различных сценариев, где важна честность модели относительно границ своих знаний.

5. Повышение качества цитирования - **Прямая применимость**: Высокая. Пользователи могут требовать от моделей предоставления цитат и использовать описанные методики для проверки их обоснованности. - **Концептуальная ценность**: Высокая. Понимание связи между утверждениями и их источниками критически важно для оценки достоверности информации. - **Потенциал для адаптации**: Высокий. Принципы качественного цитирования могут быть применены в различных контекстах, от академических исследований до фактчекинга.

Prompt:

Использование знаний из исследования TRUST-ALIGN в промптах для GPT ## Ключевые аспекты исследования для промптов

Исследование "Измерение и повышение доверия к LLM в RAG через обоснованные атрибуции и обучение отказу" предоставляет ценные знания о том, как улучшить надежность ответов языковых моделей. Основные применимые концепции:

TRUST-SCORE - комплексная метрика оценки надежности ответов **Способность отказа от ответа** при недостаточности информации **Точность цитирования** и атрибуции источников **Снижение зависимости от параметрического знания** в пользу предоставленных документов ## Пример промпта с применением знаний из исследования

[======] # Запрос для анализа медицинской информации

Контекст [Вставьте здесь релевантные медицинские документы/источники]

Инструкции для GPT: Проанализируй предоставленные медицинские документы и ответь на мой вопрос о [конкретная медицинская тема]. При формировании ответа придерживайся следующих принципов:

Если в предоставленных документах недостаточно информации для полного ответа, ЯВНО УКАЖИ ЭТО и воздержись от дополнения ответа своими знаниями.

Для каждого значимого утверждения в твоем ответе укажи конкретный источник из предоставленных документов в формате [Документ X].

Разделяй информацию на:

Факты, напрямую подтвержденные предоставленными документами (с цитированием) Выводы, которые можно логически сделать из документов (с объяснением) Области, где информация отсутствует или неполна (с явным указанием)

Не используй свои встроенные медицинские знания, если они не подтверждаются предоставленными документами.

Мой вопрос: [Ваш медицинский вопрос] [======]

Почему это работает

Данный промпт применяет принципы TRUST-ALIGN следующим образом:

Обучение отказу от ответа - явное требование указать недостаточность информации и воздержаться от использования параметрического знания

Улучшение качества цитирования - требование связывать каждое утверждение с конкретным источником

Снижение галлюцинаций - разделение информации на категории по уровню подтверждения из документов

Повышение прозрачности - структурированный формат ответа, позволяющий легко отследить источники информации

Такой подход помогает получить более надежный и проверяемый ответ от GPT, что особенно важно в критически значимых областях вроде медицины, юриспруденции или финансов.