RevisEval: Улучшение LLM в роли судьи с помощью адаптированных ответов

Дата: 2025-02-18 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2410.05193

Рейтинг: 60
Адаптивность: 75

Ключевые выводы:

Исследование предлагает новую парадигму оценки текстовых генераций под названием RevisEval, которая использует адаптированные к ответам референсы для улучшения работы LLM в качестве судьи. Основной результат: RevisEval превосходит традиционные методы оценки без референсов и с референсами в различных задачах NLG и следования инструкциям.

Объяснение метода:

Исследование RevisEval предлагает ценную концепцию оценки качества ответов LLM через создание улучшенных версий этих ответов. Хотя прямая реализация метода требует технических знаний и доступа к API, концептуальные идеи о предвзятостях моделей и эффективных стратегиях оценки могут быть адаптированы обычными пользователями. Особенно ценны выводы о том, как улучшенные версии ответов помогают выявлять недостатки в исходных ответах.

Ключевые аспекты исследования: 1. **Концепция адаптивных референсов (response-adapted references)** - исследование предлагает новую парадигму оценки генерации текста через создание "адаптированных референсов". Вместо использования фиксированных эталонных текстов, метод RevisEval создаёт референсы путём улучшения исходных ответов модели.

Двухэтапный процесс оценки - сначала LLM-ревизор улучшает исходный ответ, создавая адаптированный референс, затем этот референс используется для более точной оценки качества исходного ответа.

Применимость к традиционным метрикам - исследование демонстрирует, что адаптированные референсы значительно улучшают эффективность классических метрик оценки текста (BLEU, ROUGE, BERTScore), делая их применимыми даже для открытых задач.

Снижение предвзятости в оценках - метод RevisEval показывает меньшую позиционную предвзятость в сравнительных оценках и лучше справляется со

сложными случаями, когда модели склонны предпочитать более многословные ответы.

Альтернативная парадигма для слабых моделей - исследование предлагает использовать слабые LLM в качестве ревизоров с последующим применением классических метрик вместо прямого использования слабых моделей для оценки.

Дополнение:

Применимость в стандартном чате

Хотя авторы исследования использовали дообучение и API для реализации полноценного метода RevisEval, основные концепции могут быть адаптированы для использования в стандартном чате без специальных инструментов.

Концепции, применимые в стандартном чате:

Двухэтапная оценка ответов - пользователь может попросить модель сначала улучшить свой ответ (например, "Пожалуйста, улучши свой предыдущий ответ, сделав его более точным и информативным"), а затем проанализировать различия между исходным и улучшенным ответами.

Выявление предвзятостей - пользователь может проверить позиционную предвзятость, меняя порядок вариантов в запросе при сравнительной оценке. Также можно проверить предвзятость к многословности, предлагая модели сравнить краткий и подробный ответы.

Использование ревизии как метода улучшения - пользователь может запросить модель улучшить определенные аспекты ответа, что часто даёт лучшие результаты, чем попытка получить идеальный ответ с первого раза.

Оценка через сравнение - вместо абсолютной оценки качества, более надёжным может быть сравнительный подход, когда один ответ оценивается относительно другого (улучшенного) варианта.

Ожидаемые результаты применения:

Более критичная оценка качества ответов LLM Лучшее понимание ограничений и предвзятостей модели Более эффективное итеративное улучшение ответов Снижение влияния известных предвзятостей (к многословности, к позиции) при оценке качества Важно отметить, что метод не требует дообучения или API для базового применения - достаточно стандартного интерфейса чата, хотя эффективность будет ниже, чем у полной реализации метода RevisEval.

Анализ практической применимости: **1. Концепция адаптивных референсов** - **Прямая применимость**: Средняя. Обычные пользователи не могут напрямую применить этот метод без доступа к API или специальным инструментам. - **Концептуальная ценность**: Высокая. Понимание того, что оценка улучшается при

сравнении ответа с его улучшенной версией, может изменить подход к оценке качества ответов LLM. - **Потенциал для адаптации**: Высокий. Пользователи могут мысленно представлять "идеальную версию" ответа и использовать это для оценки качества.

2. Двухэтапный процесс оценки - **Прямая применимость**: Низкая для обычных пользователей, требует доступа к нескольким запросам к LLM. - **Концептуальная ценность**: Высокая. Демонстрирует, что разделение процесса на генерацию эталона и оценку повышает качество оценки. - **Потенциал для адаптации**: Средний. Пользователи могут запрашивать модель улучшить свой ответ, а затем сравнивать исходный и улучшенный варианты.

3. Применимость к традиционным метрикам - **Прямая применимость**: Низкая для обычных пользователей, требует технических знаний. - **Концептуальная ценность**: Средняя. Показывает, что даже простые метрики могут быть эффективны при правильном выборе референсов. - **Потенциал для адаптации**: Средний. Концепция может быть адаптирована для создания более простых инструментов оценки качества.

4. Снижение предвзятости в оценках - **Прямая применимость**: Средняя. Пользователи могут осознавать и учитывать возможные предвзятости моделей. - **Концептуальная ценность**: Высокая. Понимание типов предвзятости (позиционной, к многословности) помогает критически оценивать ответы LLM. - **Потенциал для адаптации**: Высокий. Пользователи могут использовать знание о предвзятостях для более эффективных запросов и оценки качества.

5. Альтернативная парадигма для слабых моделей - **Прямая применимость**: Низкая для обычных пользователей. - **Концептуальная ценность**: Высокая. Показывает, что генеративные способности могут быть более ценными, чем дискриминативные. - **Потенциал для адаптации**: Средний. Пользователи могут предпочитать модели, которые лучше генерируют ответы, а не те, которые лучше оценивают.

Prompt:

Использование знаний из исследования RevisEval в промптах для GPT ## Ключевые применимые концепции исследования

Исследование RevisEval предлагает метод, при котором: 1. Создаются адаптированные референсы (эталонные ответы) 2. Эти референсы используются для более точной оценки генерируемого контента 3. Даже слабые LLM могут эффективно работать как ревизоры

Пример промпта с применением знаний из RevisEval

[======] # Задача: Оценка качества ответа на технический вопрос

Контекст Я хочу, чтобы ты выступил в роли эксперта-оценщика, используя метод RevisEval. Этот метод предполагает сначала создание адаптированного референса, а затем оценку ответа относительно этого референса.

Инструкции 1. Сначала прочитай вопрос и ответ, который нужно оценить 2. Создай адаптированный референс - идеальный ответ на этот вопрос, учитывающий структуру и подход оцениваемого ответа, но исправляющий недостатки 3. Сравни оригинальный ответ с созданным тобой референсом 4. Оцени ответ по шкале от 1 до 10 по следующим критериям: - Точность (насколько информация корректна) - Полнота (насколько охвачены все аспекты вопроса) - Ясность (насколько понятно объяснение) 5. Предоставь краткое обоснование оценки, указав ключевые сильные стороны и недостатки

Вопрос для оценки [ВСТАВИТЬ ВОПРОС]

Ответ для оценки [ВСТАВИТЬ ОТВЕТ] [======]

Как это работает

Данный промпт применяет основной принцип RevisEval:

Адаптация референса к ответу: Вместо использования фиксированного эталона, GPT создает персонализированный референс, который сохраняет подход и структуру оцениваемого ответа, но исправляет его недостатки.

Двухэтапная оценка: Сначала генерируется референс, затем проводится оценка относительно этого референса, что по исследованию повышает точность оценки на 2-6%.

Многокритериальная оценка: Разделение оценки на несколько критериев (точность, полнота, ясность) соответствует рекомендации создавать тонко настроенные референсы для разных аспектов оценки.

Этот подход снижает позиционные смещения и делает оценку более объективной и стабильной, согласно выводам исследования.