Раскрытие процессов оценивания: анализ различий между LLM и человеческими оценщиками в автоматическом оценивании

Дата: 2025-02-21 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2407.18328

Рейтинг: 70
Адаптивность: 80

Ключевые выводы:

Исследование направлено на выявление различий между процессами оценивания ответов учащихся, выполняемыми большими языковыми моделями (LLM) и людьми-экспертами. Основные результаты показывают, что существует значительный разрыв в подходах к оцениванию между LLM и людьми, причем LLM часто используют 'короткие пути' вместо глубокого логического анализа, характерного для человеческого оценивания.

Объяснение метода:

Исследование раскрывает различия между оцениванием LLM и людьми, предлагая практические методы улучшения оценки. Пользователи могут запрашивать аналитические рубрики, предоставлять структурированные критерии и понимать ограничения LLM в логическом анализе. Несмотря на фокус на образовательном контексте, принципы применимы к широкому спектру задач оценивания.

Ключевые аспекты исследования: 1. **Сравнение процессов оценивания LLM и человеком**: Исследование изучает различия между тем, как LLM и люди-эксперты оценивают ответы учащихся на научные задачи.

Аналитические рубрики: Авторы побуждают LLM генерировать аналитические рубрики (наборы правил для оценки) и сравнивают их с рубриками, созданными людьми, чтобы выявить несоответствия.

Обнаружение "коротких путей": Исследование показывает, что LLM часто используют поверхностные признаки для оценки (ключевые слова), вместо следования глубоким логическим цепочкам рассуждений, как это делают люди.

Влияние примеров: Эксперименты показывают, что предоставление LLM примеров оцененных ответов учащихся может фактически снизить качество оценки, поощряя модель искать "короткие пути" вместо понимания задания.

Повышение точности: Исследование демонстрирует, что включение качественных аналитических рубрик, отражающих логику человеческой оценки, может улучшить точность оценивания LLM.

Дополнение:

Применимость методов в стандартном чате

Исследование не требует дообучения или API для применения основных концепций. Большинство методов можно адаптировать для стандартного чата с LLM:

Запрос аналитических рубрик перед оценкой Пользователь может попросить LLM создать набор критериев для оценки перед тем, как предоставить материал для оценивания Пример: "Прежде чем я покажу тебе эссе для оценки, опиши критерии, по которым ты будешь его оценивать"

Структурирование запроса на оценку

Пользователь может предоставить собственные критерии оценки Пример: "Оцени этот текст по следующим критериям: 1) логичность аргументации, 2) использование фактов, 3) стиль изложения"

Проверка процесса оценивания

Пользователь может запросить объяснение процесса оценки Пример: "Объясни, почему ты поставил такую оценку. Какие конкретные элементы текста повлияли на твое решение?"

Контроль "коротких путей"

Пользователь может проверить, не использует ли LLM поверхностные признаки Пример: "Не основывай свою оценку только на наличии ключевых слов. Оцени глубину понимания темы" Основной вывод исследования — LLM и люди могут использовать разные критерии при оценке, даже если итоговые оценки совпадают. Запрос и предоставление четких критериев оценки значительно улучшает качество оценки LLM.

Анализ практической применимости: 1. **Сравнение процессов оценивания**: - Прямая применимость: Средняя. Пользователи могут более критично относиться к оценкам, предоставляемым LLM, понимая, что модель может использовать иные критерии, чем человек. - Концептуальная ценность: Высокая. Понимание разницы между "пониманием" LLM и человеком помогает скорректировать ожидания от автоматической оценки. - Потенциал для адаптации: Средний. Пользователи могут запрашивать у LLM объяснение критериев оценки перед получением самой оценки.

Аналитические рубрики: Прямая применимость: Высокая. Пользователи могут

запрашивать у LLM создание аналитических рубрик перед выполнением задания, чтобы лучше понять критерии оценки. Концептуальная ценность: Высокая. Понимание, что LLM может генерировать рубрики, помогает использовать модели как инструмент подготовки к оценке. Потенциал для адаптации: Высокий. Техника запроса рубрик может быть применена к любой задаче, требующей оценки.

Обнаружение "коротких путей":

Прямая применимость: Средняя. Пользователи могут проверять, не использует ли LLM поверхностные признаки для оценки, запрашивая объяснение оценки. Концептуальная ценность: Высокая. Понимание ограничений LLM в логическом анализе помогает критически оценивать получаемые результаты. Потенциал для адаптации: Средний. Пользователи могут разработать стратегии формулирования запросов, требующих глубокого анализа.

Влияние примеров:

Прямая применимость: Высокая. Пользователи должны быть осторожны при предоставлении примеров LLM, понимая, что это может снизить качество оценки. Концептуальная ценность: Высокая. Понимание, что "больше примеров" не всегда означает "лучший результат", меняет подход к взаимодействию с LLM. Потенциал для адаптации: Высокий. Этот принцип применим к различным задачам взаимодействия с LLM.

Повышение точности:

Прямая применимость: Высокая. Пользователи могут предоставлять LLM качественные рубрики для улучшения точности оценки. Концептуальная ценность: Высокая. Понимание, что внешние структурированные инструкции улучшают работу LLM, применимо к различным задачам. Потенциал для адаптации: Высокий. Принцип структурирования критериев может быть применен к любой задаче оценки.

Prompt:

Использование исследования об оценивании LLM в промптах ## Ключевые выводы для создания промптов

Исследование показывает, что LLM могут эффективно оценивать ответы, но их подход отличается от человеческого. Эти знания можно использовать для создания более эффективных промптов.

Пример промпта для оценивания студенческих ответов

[======] Оцени следующий ответ студента на задание по физике.

ЗАДАНИЕ: [описание задания по физике]

ХОЛИСТИЧЕСКАЯ РУБРИКА: - Отлично (5 баллов): Полное понимание концепции,

безупречное применение формул, логичное объяснение. - Хорошо (4 балла): Хорошее понимание, небольшие ошибки в применении. - Удовлетворительно (3 балла): Базовое понимание, значительные ошибки. - Неудовлетворительно (2 балла): Серьезные концептуальные ошибки.

ПРИМЕРЫ АНАЛИТИЧЕСКИХ РУБРИК ДЛЯ ДРУГИХ ЗАДАНИЙ: 1. Задание по электричеству: - Правильное применение закона Ома (+2 балла) - Расчет сопротивления цепи (+2 балла) - Объяснение физического смысла результата (+1 балл)

ОТВЕТ СТУДЕНТА: [ответ студента]

Пожалуйста, выполни следующее: 1. Создай детальную аналитическую рубрику для данного задания с конкретными критериями оценки. 2. Оцени ответ студента по этой рубрике, анализируя логическую цепочку рассуждений, а не только наличие ключевых слов. 3. Объясни свои рассуждения для каждого пункта оценивания. 4. Укажи итоговую оценку и общее заключение. [======]

Почему этот промпт эффективен

Предоставление холистической рубрики помогает модели понять общую структуру оценивания (повышает F1-показатель).

Включение примеров аналитических рубрик из других заданий направляет модель к созданию более качественных критериев (повышает точность с 34.83% до 50.41%).

Явное требование анализировать логическую цепочку, а не искать ключевые слова, помогает избежать "коротких путей" оценивания.

Запрос на объяснение рассуждений заставляет модель использовать более глубокий анализ, как это делают люди-эксперты.

Структурированный подход (создание рубрики → оценка → объяснение → итог) следует рекомендациям исследования о сотрудничестве между LLM и экспертами.

Такой промпт значительно повышает качество оценивания LLM, приближая его к человеческому уровню экспертизы.