Автоматизированная оценка заданий с использованием больших языковых моделей: выводы из курса биоинформатики.

Дата: 2025-01-24 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2501.14499

Рейтинг: 72
Адаптивность: 85

Ключевые выводы:

Исследование оценивало эффективность использования больших языковых моделей (LLM) для автоматической проверки письменных заданий в курсе биоинформатики. Основная цель заключалась в определении, могут ли LLM заменить преподавателей при оценке и предоставлении обратной связи. Результаты показали, что при хорошо разработанных промптах LLM могут достигать точности оценивания и качества обратной связи, сопоставимых с человеческими проверяющими, причем открытые модели работают так же хорошо, как и коммерческие.

Объяснение метода:

Исследование предлагает структурированную методологию промптов (системный промпт + рубрики + примеры), которую можно адаптировать для различных задач анализа текста. Подход демонстрирует, что открытые модели могут работать не хуже коммерческих, что ценно для пользователей с ограниченным бюджетом. Хотя полная реализация требует определенных технических навыков, основные принципы доступны широкой аудитории.

Ключевые аспекты исследования: 1. **Автоматизированное оценивание письменных заданий с помощью LLM**: Исследование оценивает эффективность использования языковых моделей для автоматической проверки и оценки текстовых ответов студентов в курсе по биоинформатике.

Методология структурированных промптов: Разработан подход с использованием системных промптов, рубрик оценивания и примеров оцененных работ для обеспечения точности оценки LLM.

Сравнение моделей разных типов: Систематическое сравнение шести различных LLM (коммерческих и открытых) с человеческими оценками, показывающее, что открытые модели могут работать так же эффективно, как коммерческие.

Анализ удовлетворенности студентов: Исследование обратной связи от студентов показало, что они в целом одинаково удовлетворены оценками от LLM и от преподавателей, а в некоторых случаях даже предпочитают обратную связь от LLM.

Практические рекомендации по внедрению: Авторы предлагают конкретные рекомендации для интеграции LLM в процесс оценивания, включая структурирование рубрик, включение примеров и возможность запроса ручной проверки.

Дополнение: Методы, описанные в исследовании, в значительной степени можно применить в стандартном чате LLM без необходимости дообучения или API. Хотя в исследовании использовались различные модели (включая открытые и коммерческие), основная методология работы с LLM через структурированные промпты может быть реализована в любом стандартном чат-интерфейсе.

Ключевые концепции, которые можно применить в стандартном чате:

Структура промптов с тремя компонентами: Системный промпт с общими инструкциями Рубрики оценивания с четкими критериями Примеры оцененных работ

Балансирование рубрик и примеров: Исследование показало, что использование только рубрик приводит к более строгой оценке, а только примеров — к более снисходительной. Комбинация обоих элементов дает наиболее сбалансированные результаты. Это применимо в любом чате.

Структурированные рубрики: Разделение критериев оценки на четкие компоненты с указанием баллов за каждый критерий. Это позволяет получать более последовательные и обоснованные оценки.

Few-shot примеры: Включение 3-10 примеров оцененных работ значительно улучшает точность LLM, что можно использовать в стандартном чате для любых задач.

Запрос структурированного вывода: Указание формата ответа (например, JSON с полями "оценка", "обоснование", "обратная связь") работает в стандартном чате и помогает получать более организованные ответы.

Результаты, которые можно получить, применяя эти концепции:

Более точная и обоснованная оценка текстов различных типов (от эссе до технической документации) Более содержательная обратная связь, которая, согласно исследованию, может быть даже предпочтительнее человеческой Последовательность в оценке, сравнимая с человеческой Возможность улучшить точность анализа даже с помощью меньших моделей при правильной структуре промптов Интересно, что исследование показало: для правильно настроенного

промпта с рубриками и примерами, открытые модели работают почти так же хорошо, как коммерческие, что особенно ценно для пользователей, которые используют только бесплатные версии LLM.

Анализ практической применимости: **1. Автоматизированное оценивание с помощью LLM:** - **Прямая применимость**: Высокая для образовательных учреждений и преподавателей; средняя для обычных пользователей, которые могут адаптировать подход для оценки текстов, но без доступа к системам учета оценок. - **Концептуальная ценность**: Значительная, демонстрирует способность LLM к объективной оценке текста при правильном структурировании задачи. - **Потенциал для адаптации**: Высокий, методология может быть адаптирована для проверки и оценки различных типов текстов — от эссе до деловой документации.

2. Методология структурированных промптов: - **Прямая применимость**: Очень высокая, пользователи могут напрямую использовать описанную структуру промптов (системный промпт + рубрики + примеры) для получения более точных результатов. - **Концептуальная ценность**: Существенная, показывает важность комбинирования критериев оценки и примеров для улучшения точности LLM. - **Потенциал для адаптации**: Высокий, подход может быть применен к широкому спектру задач, требующих оценки или анализа текста.

3. Сравнение моделей разных типов: - **Прямая применимость**: Средняя, результаты сравнения помогают выбрать подходящую модель, но требуют технических знаний для реализации. - **Концептуальная ценность**: Высокая, демонстрирует, что квантованные открытые модели могут работать на уровне коммерческих, что важно для понимания возможностей LLM. - **Потенциал для адаптации**: Средний, зависит от технических возможностей пользователя и доступа к вычислительным ресурсам.

4. Анализ удовлетворенности студентов: - **Прямая применимость**: Низкая для обычных пользователей, более значима для преподавателей. - **Концептуальная ценность**: Средняя, подтверждает, что LLM могут предоставлять обратную связь, воспринимаемую так же положительно, как и человеческая. - **Потенциал для адаптации**: Средний, результаты могут быть использованы для улучшения обратной связи в различных контекстах.

5. Практические рекомендации по внедрению: - **Прямая применимость**: Высокая, рекомендации могут быть непосредственно использованы при настройке LLM для задач оценивания. - **Концептуальная ценность**: Высокая, предоставляет четкую структуру для эффективного использования LLM. - **Потенциал для адаптации**: Очень высокий, рекомендации универсальны и могут быть адаптированы для различных контекстов взаимодействия с LLM.

Prompt:

Применение исследования о LLM-оценке заданий в промптах ## Ключевые аспекты из исследования для использования в промптах

Исследование показывает, что LLM могут эффективно оценивать студенческие работы при правильной структуре промптов. Особенно важны:

Структурированные рубрики оценивания Включение примеров оцененных работ Четкие критерии для различных уровней выполнения заданий ## Пример эффективного промпта для оценки студенческих работ

[======] # Задание для оценки студенческой работы по биоинформатике

Контекст задания Ты - ассистент преподавателя курса биоинформатики. Тебе нужно оценить ответ студента на вопрос о методах выравнивания последовательностей.

Вопрос для студента "Опишите алгоритм Нидлмана-Вунша и объясните, как он используется для глобального выравнивания последовательностей."

Рубрика оценивания (по 5-балльной шкале) 1. Понимание принципа алгоритма (0-2 балла) 2. Описание матрицы замен и штрафов за пробелы (0-1 балл) 3. Объяснение процесса обратного прослеживания (0-1 балл) 4. Приведение примера применения (0-1 балл)

Примеры оцененных работ ### Пример отличного ответа (5 баллов): [Вставить образец отличного ответа]

Пример удовлетворительного ответа (3 балла): [Вставить образец среднего ответа]

Формат обратной связи 1. Общая оценка (X/5 баллов) 2. Краткое обоснование оценки 3. Конкретные комментарии по каждому пункту рубрики 4. Рекомендации по улучшению

Ответ студента для оценки: [Вставить ответ студента]

Оцени работу и предоставь структурированную обратную связь согласно указанному формату. [======]

Почему этот промпт работает в соответствии с исследованием

Структурированная рубрика - разбивает оценивание на конкретные компоненты с четкими критериями, что согласно исследованию повышает точность оценки до 85-90%

Примеры оцененных работ - исследование показало, что включение образцов помогает LLM лучше понять ожидания и стиль оценивания

Четкий формат обратной связи - структурированный шаблон для ответа, который

исследование определило как более предпочтительный для студентов

Контекст и специфика задания - детальное описание помогает модели точнее понять предметную область

Такой подход к составлению промптов позволяет достичь качества оценивания, сравнимого с человеческим, как показало исследование, особенно при использовании более крупных моделей (Llama-405Bq4, GPT-4o и подобных).