За пределами корреляции: Влияние человеческой неопределенности на измерение эффективности автоматической оценки и LLM как судьи

Дата: 2025-01-27 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2410.03775

Рейтинг: 62
Адаптивность: 75

Ключевые выводы:

Исследование направлено на анализ эффективности автоматической оценки генеративных моделей, включая LLM-as-a-Judge. Основной вывод: корреляционные метрики могут создавать иллюзию, что автоматическая оценка приближается к человеческой, когда в данных высока доля неопределенности в человеческих оценках, но при увеличении согласованности человеческих оценок корреляция между машинными и человеческими метками значительно падает.

Объяснение метода:

Исследование раскрывает важные ограничения LLM как судей и предлагает методы для более точной оценки. Ключевая ценность — понимание влияния неопределенности в человеческих оценках на работу LLM. Стратификация задач по уровню определенности и многокритериальный подход к оценке имеют практическую ценность, однако технические методы требуют значительной адаптации для широкой аудитории.

Ключевые аспекты исследования: 1. **Анализ ограничений корреляционных метрик**: Исследование показывает, что традиционные корреляционные метрики (Krippendorff's α, Cohen's κ и др.) могут создавать ложное впечатление о качестве автоматической оценки LLM, особенно когда доля образцов с неопределенностью в человеческих оценках высока.

Стратификация данных по неопределенности: Авторы предлагают стратифицировать данные по уровню согласованности человеческих оценок, что позволяет выявить истинные расхождения между человеческими и автоматическими оценками.

Новая метрика оценки согласованности: Предложена метрика "binned Jensen-Shannon Divergence" (JSb), которая лучше учитывает вариативность

человеческих восприятий, не полагаясь на единственную "золотую" метку.

Техники визуализации: Разработаны методы визуализации ("perception charts"), которые наглядно демонстрируют различия между человеческими и машинными оценками, помогая интерпретировать корреляционные метрики.

Многометрический подход: Авторы рекомендуют использовать несколько метрик из разных семейств для комплексной оценки эффективности автоматической оценки.

Дополнение:

Применимость методов исследования в стандартном чате

Данное исследование не требует дообучения или специального API для применения большинства его концептуальных выводов. Хотя авторы использовали расширенные техники для анализа данных, основные принципы можно адаптировать для работы в стандартном чате:

Стратификация запросов по уровню определенности Пользователи могут разделять свои запросы на "объективные" (с однозначными ответами) и "субъективные" (допускающие вариативность) При оценке ответов LLM можно учитывать, что в субъективных задачах модель может давать ответы, отличающиеся от ожидаемых, даже если она работает корректно

Многокритериальная оценка

Вместо оценки ответа LLM по единственному критерию, пользователи могут разработать несколько критериев оценки (точность, полнота, творческий подход и т.д.) Это помогает получить более комплексное представление о качестве ответа

Учет вариативности восприятия

Понимание, что для многих задач не существует единственно правильного ответа, помогает формулировать более гибкие запросы Можно запрашивать у LLM несколько вариантов ответа с обоснованием, а не единственное решение

Корректировка ожиданий

Исследование показывает, что LLM хуже справляются с задачами, где люди демонстрируют высокое согласие Пользователи могут скорректировать свои ожидания, понимая, что в задачах с высокой определенностью может потребоваться более тщательная формулировка запроса

Улучшенные промпты для LLM-судей

При использовании LLM для оценки других ответов (LLM-as-a-Judge) пользователи могут включать в промпт указания учитывать возможную вариативность ответов для

субъективных задач Можно запрашивать не только бинарную оценку (правильно/неправильно), но и оценку с учетом распределения возможных ответов Применение этих концепций может значительно повысить эффективность взаимодействия с LLM в стандартном чате без необходимости в специальных технических инструментах или API.

Анализ практической применимости: 1. **Анализ ограничений корреляционных метрик** - Прямая применимость: Высокая. Пользователи могут осознать, что высокая корреляция между LLM-оценками и человеческими может быть иллюзорной из-за неопределенности в человеческих оценках. Это поможет избежать переоценки возможностей LLM как судьи. - Концептуальная ценность: Очень высокая. Понимание того, что LLM показывают худшую корреляцию с человеческими оценками в случаях, когда люди достигают высокого согласия, критически важно для реалистичной оценки возможностей моделей. - Потенциал для адаптации: Средний. Пользователи могут адаптировать этот принцип, уделяя больше внимания случаям, где у них есть высокая уверенность в своей оценке.

Стратификация данных по неопределенности Прямая применимость: Средняя. Рядовым пользователям сложно самостоятельно реализовать стратификацию, но они могут применять принцип разделения задач по уровню определенности. Концептуальная ценность: Высокая. Понимание необходимости разделять задачи на те, где есть четкие критерии и где оценка субъективна, помогает формулировать более точные запросы к LLM. Потенциал для адаптации: Высокий. Пользователи могут внедрить практику указания уровня уверенности в своих запросах и оценках.

Новая метрика оценки согласованности (JSb)

Прямая применимость: Низкая. Сложная для внедрения рядовыми пользователями. Концептуальная ценность: Средняя. Понимание, что не всегда существует единственно правильный ответ, важно для адекватной оценки результатов LLM. Потенциал для адаптации: Средний. Пользователи могут применять принцип рассмотрения распределения возможных ответов, а не поиска единственно правильного.

Техники визуализации

Прямая применимость: Средняя. Создание подобных визуализаций требует технических навыков, но принцип сравнения распределений ответов доступен большинству пользователей. Концептуальная ценность: Высокая. Визуализация помогает лучше понять, как LLM "воспринимает" задачу по сравнению с людьми. Потенциал для адаптации: Высокий. Пользователи могут разработать простые способы визуального сравнения ответов LLM с ожидаемыми результатами.

Многометрический подход

Прямая применимость: Средняя. Использование нескольких критериев оценки доступно большинству пользователей. Концептуальная ценность: Высокая. Понимание, что нет единственной метрики для оценки качества ответов LLM,

помогает формировать более комплексное представление о возможностях моделей. Потенциал для адаптации: Высокий. Пользователи могут разработать собственные многокритериальные системы оценки для своих задач.

Prompt:

Применение исследования о человеческой неопределенности в промтах для GPT ## Ключевые аспекты исследования для промптинга

Исследование показывает, что традиционные метрики корреляции могут создавать иллюзию эффективности автоматической оценки, особенно когда в данных высока доля неопределенности в человеческих оценках. Это знание можно эффективно применить при создании промптов для GPT.

Пример промпта с учетом исследования

[======] Оцени качество следующего текстового резюме по шкале от 1 до 5, где: 1 - очень плохо, 5 - отлично.

При оценке учитывай следующие факторы: - Информативность (полнота передачи ключевой информации) - Связность (логическая структура текста) - Читабельность (ясность изложения)

Важно: Вместо единственной оценки, предоставь распределение вероятностей для каждой оценки (от 1 до 5) по каждому критерию, отражая возможную вариативность человеческих мнений. Затем объясни свое решение и укажи, для каких аспектов текста характерна наибольшая неопределенность в оценке.

Текст для оценки: [Текст резюме] [======]

Объяснение эффективности

Данный промпт применяет знания из исследования следующим образом:

Учет неопределенности оценок: Вместо единственной оценки запрашивается распределение вероятностей, что отражает реальную вариативность человеческих суждений.

Стратификация по критериям: Разделение оценки на конкретные критерии позволяет выявить области, где неопределенность выше или ниже.

Явное обозначение неопределенности: Требование указать аспекты с наибольшей неопределенностью помогает избежать иллюзии точности там, где ее объективно не может быть.

Многомерность оценки: Использование нескольких критериев вместо одной агрегированной оценки соответствует рекомендациям исследования.

Такой подход к промптингу поможет получить более реалистичные и информативные оценки от GPT, избегая ловушек, связанных с упрощенным пониманием корреляции между машинными и человеческими оценками.