Максимизация сигнала в соответствии предпочтений человека и модели

Дата: 2025-03-06 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2503.04910

Рейтинг: 62
Адаптивность: 75

Ключевые выводы:

Исследование направлено на разработку методологии для интеграции человеческих предпочтений в обучение и оценку LLM-моделей. Основной вывод: в случаях, когда конечные пользователи должны соглашаться с решениями моделей (например, при обнаружении токсичности или извлечении ключевых моментов), модели должны обучаться и оцениваться на данных, отражающих предпочтения этих пользователей.

Объяснение метода:

Исследование предлагает ценную концептуальную основу для понимания субъективности в ответах LLM, разделяя "шум" (ошибки) от "сигнала" (значимых разногласий). Высокая применимость для критической оценки ответов и формулировки запросов, но требует адаптации технических методов для широкого использования. Особенно полезно понимание типов субъективности задач и их влияния на ожидания от LLM.

Ключевые аспекты исследования: 1. **Разграничение шума и сигнала в задачах оценки**: исследование предлагает методологию различения "шума" (случайных ошибок) от "сигнала" (осмысленных разногласий) в субъективных оценках человеческих предпочтений для LLM.

Онтология субъективности: авторы классифицируют задачи оценки на три типа по шкале субъективности: явное содержание (объективное), скрытые паттерны (полусубъективное) и проективное содержание (полностью субъективное).

Методология сбора данных: исследование предлагает конкретные подходы к сбору качественных данных о человеческих предпочтениях, включая размер выборки, методы выборки и анализ межэкспертной согласованности.

Практический кейс: авторы демонстрируют применение своей методологии на примере оценки двух моделей-классификаторов для функции блокировки нежелательного контента, используя оценки людей для согласования поведения

модели с предпочтениями пользователей.

Количественные методы оценки согласованности между аннотаторами с использованием статистических инструментов (коэффициенты каппа Коэна, альфа Криппендорфа).

Дополнение: Действительно ли для работы методов этого исследования требуется дообучение или API? Или методы и подходы можно применить в стандартном чате?

Данное исследование **не требует** дообучения моделей или специального API для применения большинства концепций. Ключевые концепции и подходы можно адаптировать для использования в стандартном чате:

Классификация субъективности запросов: Пользователи могут самостоятельно оценивать, к какому типу относится их запрос (объективный, полусубъективный, полностью субъективный) Это позволяет корректировать ожидания от ответа и формулировку запроса

Разделение шума и сигнала:

При получении противоречивых ответов от LLM пользователи могут определять, вызваны ли противоречия ошибками или субъективностью вопроса Можно задавать уточняющие вопросы для проверки согласованности ответов

Многократные запросы для субъективных тем:

Для важных субъективных вопросов можно использовать несколько переформулировок запроса Различные ответы можно интерпретировать как сигнал о разнообразии мнений, а не как ошибку

Явное запрашивание разных точек зрения:

Для субъективных вопросов можно явно просить модель представить разные перспективы Это реализует идею исследования о ценности разнообразия мнений Результаты применения этих подходов: - Более реалистичные ожидания от взаимодействия с LLM - Лучшее понимание ограничений моделей в субъективных вопросах - Более информированное использование ответов в зависимости от типа задачи - Повышение критического мышления при оценке ответов LLM

Хотя исследователи использовали статистические методы и масштабные опросы, основные концепции применимы и в индивидуальном взаимодействии с LLM.

Анализ практической применимости: 1. **Разграничение шума и сигнала** - **Прямая применимость**: Высокая. Пользователи могут применить предложенную классификацию для понимания, когда разногласия в оценках являются информативными, а когда просто ошибкой. - **Концептуальная ценность**: Очень высокая. Понимание разницы между шумом и сигналом в оценках помогает

формировать более эффективные запросы к LLM и интерпретировать вариативность ответов. - **Потенциал для адаптации**: Высокий. Концепцию можно использовать при создании пользовательских систем обратной связи и оценки ответов LLM.

Онтология субъективности Прямая применимость: Средняя. Классификация типов задач помогает пользователям осознать, в каких случаях стоит ожидать разных ответов от LLM. **Концептуальная ценность**: Высокая. Понимание спектра субъективности задач улучшает понимание возможностей и ограничений LLM в разных контекстах. **Потенциал для адаптации**: Высокий. Пользователи могут адаптировать свои ожидания и формулировки запросов в зависимости от типа задачи.

Методология сбора данных

Прямая применимость: Низкая для обычных пользователей, высокая для команд, работающих с LLM. **Концептуальная ценность**: Средняя. Понимание важности размера выборки и методов выборки помогает критически оценивать утверждения о возможностях LLM. **Потенциал для адаптации**: Средний. Методы можно упростить для использования в пользовательских опросах и оценочных системах.

Практический кейс

Прямая применимость: Средняя. Кейс демонстрирует конкретный пример применения методологии, который можно адаптировать для других задач. **Концептуальная ценность**: Высокая. Показывает, как человеческие оценки могут влиять на выбор модели. **Потенциал для адаптации**: Высокий. Подход можно масштабировать для различных сценариев использования LLM.

Количественные методы

Прямая применимость: Низкая для обычных пользователей, высокая для команд, работающих с LLM. **Концептуальная ценность**: Средняя. Понимание статистических методов оценки согласованности улучшает критическое мышление. **Потенциал для адаптации**: Средний. Методы могут быть упрощены для базовой оценки согласованности между пользователями.

Prompt:

Применение исследования о человеческих предпочтениях в промптах для GPT ## Ключевая идея исследования

Исследование показывает, что при работе с LLM важно учитывать субъективность задачи и согласованность модели с предпочтениями пользователей, особенно когда задачи не имеют однозначно правильных ответов.

Пример промпта с учетом выводов исследования

[======] Я хочу, чтобы ты оценил следующий текст на предмет токсичности, учитывая, что это субъективная задача с проективным содержанием (latent projective content).

Текст для анализа: [ВСТАВИТЬ ТЕКСТ]

Вместо однозначного вердикта "токсичный/нетоксичный", пожалуйста: 1. Оцени вероятность того, что разные группы людей могут счесть текст токсичным (например, "70% вероятность для группы X, 30% для группы Y") 2. Объясни различные точки зрения, которые могут существовать по поводу этого текста 3. Предложи несколько вариантов смягчения текста с разной степенью изменения исходного сообщения

Это поможет мне лучше понять спектр возможных реакций и принять более информированное решение. [======]

Объяснение эффективности промпта

Данный промпт применяет знания из исследования следующим образом:

Признает субъективность задачи - вместо поиска единственно верного ответа, признает, что оценка токсичности относится к проективному содержанию

Работает с распределением мнений - запрашивает вероятностную оценку для разных групп людей, что соответствует рекомендации использовать мягкие метрики вместо жестких

Сохраняет разнообразие интерпретаций - просит объяснить различные точки зрения, сохраняя "сигнал" в разногласиях между возможными аннотаторами

Учитывает предпочтения конечных пользователей - предлагает варианты решений, которые могут соответствовать разным ожиданиям пользователей

Такой подход к составлению промптов делает взаимодействие с GPT более информативным и полезным для субъективных задач, где важно учитывать разнообразие человеческих предпочтений.