Оценка воспринимаемой уверенности для аннотирования данных с помощью Zero-Shot LLM

Дата: 2025-02-10 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2502.07186

Рейтинг: 78
Адаптивность: 85

Ключевые выводы:

Исследование представляет новую технику Perceived Confidence Scoring (PCS) для оценки уверенности LLM в задачах классификации текста. Основная цель - повысить точность и надежность аннотаций, создаваемых LLM в режиме zero-shot. Главные результаты показывают, что PCS значительно улучшает точность классификации по сравнению с традиционными методами, такими как голосование большинством, в среднем на 4.96-10.52% для отдельных моделей и на 7.75% при использовании нескольких моделей.

Объяснение метода:

Исследование предлагает практичный метод оценки надежности LLM через проверку согласованности ответов на переформулированные запросы. Основные концепции (метаморфические отношения) легко применимы обычными пользователями через простые промпты, значительно повышая точность классификации. Хотя полная реализация оптимизации весов требует технических знаний, базовый подход доступен всем.

Ключевые аспекты исследования: 1. **Методика Perceived Confidence Score (PCS)** - новый подход для оценки уверенности LLM при классификации текста, основанный на анализе согласованности ответов модели при семантически эквивалентных, но текстуально различных вариантах входных данных.

Метаморфические отношения (MR) - три типа трансформаций текста: (1) преобразование активного/пассивного залога, (2) двойное отрицание, (3) замена синонимов. Эти преобразования сохраняют смысл текста, но меняют его форму.

Perceived Differential Evolution (PDE) - алгоритм оптимизации, который определяет оптимальные веса для метаморфических отношений и выходных данных LLM, повышая точность классификации.

Применение для нескольких LLM - метод PCS может использоваться как для одной модели, так и для комбинации нескольких моделей, превосходя

традиционные методы вроде голосования большинством.

Эмпирическая валидация - исследование показало значительное повышение точности классификации на четырех наборах данных с использованием трех разных LLM: Llama 3, Mistral и Gemma.

Дополнение: Методы этого исследования **не требуют дообучения или специального API** для базового применения. Основные концепции могут быть реализованы в стандартном чате с LLM. Ученые использовали продвинутые техники (PDE) для оптимизации и количественной оценки, но сама методология работает и без них.

Применимые в стандартном чате концепции:

Базовая проверка согласованности - пользователь может задать один и тот же вопрос несколькими способами и сравнить ответы. Если они согласованы, вероятно, модель более уверена.

Метаморфические преобразования текста - три типа трансформаций из исследования могут быть легко применены:

Изменение активного/пассивного залога: "Как Apple повлияла на рынок смартфонов?" → "Как рынок смартфонов был изменен компанией Apple?" Двойное отрицание: "Полезно ли есть овощи?" → "Не вредно ли не есть овощи?" Замена синонимов: "Как создать эффективный бизнес-план?" → "Как разработать результативную бизнес-стратегию?"

Простое взвешенное голосование - если модель дает одинаковый ответ на большинство вариаций вопроса, этот ответ, вероятно, более надежен.

Ожидаемые результаты:

Повышение качества принятия решений - пользователи смогут выявлять ситуации, когда LLM "не уверена" в своем ответе, и соответственно корректировать свое доверие к информации.

Выявление слабых мест модели - некоторые типы переформулировок могут выявлять области, где модель особенно чувствительна к формулировкам.

Более надежная классификация - при задачах, требующих категоризации (например, определение тональности текста, классификация содержания), этот подход может значительно повысить точность без технических сложностей.

Даже без сложной оптимизации весов, простое применение этих концепций может повысить надежность работы с LLM на 5-10%, что согласуется с результатами исследования.

Анализ практической применимости: 1. **Методика Perceived Confidence Score**

(PCS) - Прямая применимость: Высокая. Пользователи могут легко адаптировать подход, запрашивая LLM классифицировать исходный текст и его вариации, затем оценивая согласованность ответов. - Концептуальная ценность: Очень высокая. Метод предлагает понимание того, как уверенность LLM проявляется в согласованности ответов на разные формулировки, что помогает пользователям лучше оценивать надежность ответов. - Потенциал для адаптации: Высокий. Принцип проверки согласованности можно применить к широкому спектру задач, не ограничиваясь классификацией.

Метаморфические отношения (MR) Прямая применимость: Высокая. Предложенные трансформации просты и могут быть реализованы обычными пользователями через промпты. Концептуальная ценность: Высокая. Понимание того, как различные переформулировки влияют на ответы LLM, позволяет пользователям более критично оценивать результаты. Потенциал для адаптации: Очень высокий. Пользователи могут разработать собственные типы трансформаций для конкретных задач.

Perceived Differential Evolution (PDE)

Прямая применимость: Средняя. Требует технических знаний для реализации, но концепция оптимизации весов может быть упрощена. Концептуальная ценность: Средняя. Помогает понять, что разные типы переформулировок имеют разную эффективность. Потенциал для адаптации: Средний. Можно использовать более простые методы взвешивания результатов.

Применение для нескольких LLM

Прямая применимость: Средняя. Требует доступа к нескольким LLM, что может быть сложно для обычных пользователей. Концептуальная ценность: Высокая. Демонстрирует преимущества комбинирования нескольких моделей над голосованием большинством. Потенциал для адаптации: Средний. Концепцию можно применить к разным доступным моделям.

Эмпирическая валидация

Прямая применимость: Низкая. Результаты сами по себе не применимы напрямую. Концептуальная ценность: Высокая. Подтверждает эффективность подхода и мотивирует к его применению. Потенциал для адаптации: Высокий. Методология тестирования может быть адаптирована для оценки эффективности в других контекстах.

Prompt:

Использование метода Perceived Confidence Scoring (PCS) в промптах для GPT ## Суть метода PCS Метод PCS повышает точность классификации текста путем: 1. Создания нескольких вариаций исходного текста (метаморфические отношения) 2. Оценки согласованности ответов модели по этим вариациям 3. Определения уровня "воспринимаемой уверенности" модели

Пример промпта с применением PCS

[======] Я хочу, чтобы ты классифицировал следующий текст по тональности (позитивный, негативный или нейтральный). Для повышения точности я предоставлю одну и ту же информацию в 3 различных формулировках. Пожалуйста:

Классифицируй каждую версию отдельно Проанализируй согласованность своих ответов Укажи свой финальный ответ с объяснением уровня уверенности Если твои классификации для разных версий различаются, объясни почему Версия 1: "Этот новый телефон имеет отличную камеру, но батарея разряжается слишком быстро." Версия 2: "Камера в этом новом телефоне превосходная, однако батарея держится недолго." Версия 3: "Несмотря на высокое качество фотографий, которые делает новый телефон, его аккумулятор быстро садится." [======]

Как это работает

Метаморфические отношения: Создаются семантически эквивалентные, но текстуально различные версии входного текста (в примере - три перефразированные версии отзыва о телефоне).

Оценка согласованности: GPT анализирует все версии и сравнивает свои ответы. Если ответы совпадают, это указывает на высокую уверенность модели.

Улучшение точности: Согласно исследованию, такой подход повышает точность классификации на 5-10% по сравнению с простым запросом.

Прозрачность: Метод делает процесс принятия решений более интерпретируемым, поскольку модель объясняет свою уверенность на основе согласованности ответов.

Этот подход особенно полезен для задач классификации текста, анализа тональности, обнаружения фейковых новостей и других задач, где важна точность и надежность результатов.