Оценка надежности самообъяснений в больших языковых моделях

Дата: 2025-01-31 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2407.14487

Рейтинг: 72
Адаптивность: 85

Ключевые выводы:

Исследование направлено на оценку надежности самообъяснений, генерируемых большими языковыми моделями (LLM) при запросе объяснить свой предыдущий вывод. Основные результаты показывают, что хотя самообъяснения LLM могут коррелировать с человеческими оценками, они не всегда точно отражают внутренний процесс принятия решений модели. Однако этот разрыв можно преодолеть с помощью контрфактических самообъяснений, которые могут быть достоверными, информативными и легко проверяемыми.

Объяснение метода:

Исследование предлагает практические методы получения самообъяснений от LLM через простые промпты. Контрфактуальные объяснения особенно полезны, так как позволяют понять ключевые факторы принятия решений и легко проверяются. Эти подходы не требуют технических знаний и могут применяться с любым LLM. Однако для максимальной эффективности требуется адаптация промптов под конкретные задачи.

Ключевые аспекты исследования: 1. Исследование оценивает надежность самообъяснений (self-explanations), генерируемых LLM, когда модель объясняет свои собственные выводы. 2. Рассматриваются два типа самообъяснений: экстрактивные (extractive) - выделение ключевых фраз, повлиявших на решение, и контрфактуальные (counterfactual) - версии текста, меняющие решение модели при минимальных изменениях. 3. Проводится сравнение самообъяснений с традиционными методами объяснимости (внимание и градиенты) и с оценками людей. 4. Исследование показывает, что самообъяснения хорошо коррелируют с человеческими оценками, но не всегда точно отражают внутренние процессы модели. 5. Контрфактуальные объяснения показывают высокую верность (faithfulness) и могут служить альтернативой традиционным методам объяснимости.

Дополнение: Исследование не требует дообучения или специального API для применения основных методов. Ключевые концепции, которые можно использовать в стандартном чате:

Экстрактивные самообъяснения - можно просто попросить модель объяснить свое решение или указать наиболее важные фразы, повлиявшие на её вывод. Пример промпта: "Какие фразы/слова были наиболее важными для твоего вывода? Укажи только эти фразы."

Контрфактуальные объяснения - можно попросить модель изменить минимальное количество слов в исходном тексте, чтобы получить противоположное решение. Пример промпта: "Предоставь версию этого текста, которая изменит твою оценку на противоположную, меняя как можно меньше слов. Отвечай только измененной версией."

Проверка контрфактуальных объяснений - можно предложить модели оценить измененную версию текста, чтобы проверить, действительно ли изменилось решение.

Хотя исследователи использовали градиентные и основанные на внимании методы для анализа, эти технические подходы нужны были только для исследовательских целей сравнения с самообъяснениями, а не для самого получения объяснений.

Применяя эти концепции, пользователи могут: - Получать более прозрачное понимание решений LLM - Выявлять ключевые факторы, влияющие на классификацию - Проверять последовательность модели - Улучшать формулировки запросов для получения более точных ответов

Исследование показывает, что самообъяснения часто хорошо коррелируют с человеческой интуицией, что делает их полезным инструментом для обычных пользователей.

Анализ практической применимости: **Самообъяснения LLM и их корреляция с человеческими оценками:** - Прямая применимость: Пользователи могут запрашивать у LLM объяснения принятых решений, получая ответы, которые хорошо согласуются с человеческим пониманием. - Концептуальная ценность: Понимание того, что LLM способны генерировать объяснения, которые кажутся разумными людям, даже если они не точно отражают внутренние процессы модели. - Потенциал для адаптации: Метод не требует специального обучения и может быть применен к любой задаче, где LLM делает определенный выбор.

Контрфактуальные объяснения: - Прямая применимость: Пользователи могут запрашивать у LLM версию текста, которая приведет к противоположному решению, что помогает понять ключевые факторы. - Концептуальная ценность: Высокая - дает понимание, какие именно элементы текста влияют на решение модели. - Потенциал для адаптации: Метод легко адаптируется для различных задач классификации и может быть улучшен с помощью более точных промптов.

Сравнение с традиционными методами объяснимости: - Прямая применимость: Ограниченная - требуется доступ к внутренним параметрам модели. -

Концептуальная ценность: Средняя - показывает, что привычные методы объяснимости не всегда согласуются с человеческой интуицией. - Потенциал для адаптации: Низкий - требуются специальные технические знания и доступ к модели.

Методология проверки надежности объяснений: - Прямая применимость: Пользователи могут проверить контрфактуальные объяснения, просто передав их обратно в модель. - Концептуальная ценность: Высокая - предлагает способ проверки достоверности объяснений. - Потенциал для адаптации: Средний - методология может быть адаптирована, но требует понимания принципов оценки объяснений.

Выводы о необходимости адаптации промптов: - Прямая применимость: Высокая - пользователи могут улучшать качество объяснений, адаптируя промпты. - Концептуальная ценность: Высокая - понимание важности формулировки запроса для получения качественных объяснений. - Потенциал для адаптации: Высокий - принципы формулировки эффективных промптов могут применяться широко.

Prompt:

Применение знаний об объяснениях LLM в промптах ## Ключевые выводы исследования

Исследование показывает, что **контрфактические самообъяснения** (объяснения, показывающие, как изменение входных данных влияет на результат) более достоверны и проверяемы, чем простые экстрактивные объяснения (указание на важные части текста).

Пример эффективного промпта

[======] Я хочу, чтобы ты проанализировал следующий текст отзыва и определил его тональность (позитивная/негативная):

[ТЕКСТ ОТЗЫВА]

После анализа, пожалуйста: 1. Укажи свое решение о тональности 2. Предоставь контрфактическую версию текста, которая бы изменила тональность на противоположную, изменив как можно меньше слов в оригинале 3. Объясни, почему именно эти изменения повлияли на оценку тональности

Это поможет мне лучше понять твой процесс принятия решений и проверить достоверность твоего анализа. [======]

Почему этот промпт работает

Использует контрфактические объяснения — согласно исследованию, они имеют высокую достоверность (до 95% для больших моделей)

Позволяет проверить объяснение — изменённую версию можно снова подать на вход модели, чтобы убедиться, что она действительно меняет предсказание

Адаптирован под конкретную задачу — промпт указывает на необходимость минимальных изменений и конкретный класс, на который нужно изменить оценку

Обеспечивает прозрачность — требует объяснения причин, по которым изменения влияют на результат

Практическое применение

Такой подход к промптам можно использовать в различных задачах классификации, от анализа тональности до оценки безопасности контента, когда важно не только получить результат, но и понять, почему модель приняла то или иное решение, с возможностью проверить эти объяснения на достоверность.