Управляемые подсказками внутренние состояния для обнаружения галлюцинаций в больших языковых моделях

Дата: 2025-02-27 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2411.04847

Рейтинг: 70
Адаптивность: 80

Ключевые выводы:

Исследование направлено на улучшение обнаружения галлюцинаций в больших языковых моделях (LLM) с помощью нового фреймворка PRISM (Prompt-guided Internal States for hallucination detection of LLMs). Основная цель - повысить кросс-доменную производительность детекторов галлюцинаций, обученных только на данных одного домена. Главный результат: использование специальных промптов значительно улучшает структуру внутренних состояний LLM, связанную с достоверностью текста, делая её более заметной и согласованной между разными доменами, что повышает точность обнаружения галлюцинаций.

Объяснение метода:

Исследование предлагает метод обнаружения галлюцинаций в LLM через управляемые промптами внутренние состояния. Хотя технические аспекты недоступны обычным пользователям, концепция использования специальных формулировок вопросов для проверки достоверности информации имеет высокую практическую ценность. Предложенные промпты могут быть непосредственно использованы широкой аудиторией.

Ключевые аспекты исследования: 1. **Метод обнаружения галлюцинаций в LLM**: Исследование представляет фреймворк PRISM (Prompt-guided Internal States for hallucination detection of LLMs), который использует специальные промпты для улучшения обнаружения галлюцинаций в языковых моделях.

Управляемые промптами внутренние состояния: Авторы показывают, что правильно подобранные промпты могут изменять внутренние состояния LLM таким образом, что структуры, связанные с правдивостью текста, становятся более выраженными и согласованными между разными доменами.

Улучшение кросс-доменной производительности: Исследование демонстрирует, что предложенный подход значительно улучшает способность детекторов

галлюцинаций, обученных на данных одного домена, работать с текстами из других доменов.

Метод выбора промптов: Предложен метод генерации и выбора эффективных промптов для задачи обнаружения галлюцинаций с помощью анализа отношения дисперсий.

Интеграция с существующими методами: PRISM может быть интегрирован с различными существующими методами обнаружения галлюцинаций, значительно улучшая их производительность.

Дополнение: Для реализации полного метода из исследования действительно требуется доступ к внутренним состояниям модели, что недоступно в стандартном чате. Однако ключевые концепции и подходы можно адаптировать для использования в обычном взаимодействии с LLM без необходимости в дообучении или API.

Основные адаптируемые концепции:

Использование специальных промптов для проверки достоверности. Исследование предлагает 10 различных формулировок промптов, которые эффективно помогают модели "осознать", когда она генерирует потенциально недостоверную информацию. Эти промпты можно использовать напрямую: "Является ли утверждение '[утверждение]' точным отражением истины?" "Можно ли подтвердить, что '[утверждение]' является правдой?" "Было бы правильно сказать, что '[утверждение]' является точным?"

Проверка через переформулировку. Пользователи могут переформулировать полученную информацию и попросить модель подтвердить её достоверность, используя предложенные в исследовании формулировки.

Мета-вопросы о достоверности. Можно задавать модели вопросы о её уверенности в предоставляемой информации, что концептуально близко к анализу внутренних состояний.

Кросс-доменное обобщение. Исследование показывает, что одни и те же промпты эффективны в разных предметных областях, что означает, что пользователи могут применять одинаковые стратегии проверки независимо от темы.

Ожидаемые результаты от применения этих концепций: - Повышение точности получаемой информации - Снижение риска принятия недостоверной информации - Развитие более критического подхода к взаимодействию с LLM - Улучшение способности отличать фактическую информацию от предположений или неточностей

Важно отметить, что хотя полный метод с анализом внутренних состояний недоступен, сама идея использования специальных промптов для улучшения распознавания правдивости информации является ценной и применимой частью

исследования.

Анализ практической применимости: 1. **Метод обнаружения галлюцинаций в LLM** - Прямая применимость: Средняя. Обычные пользователи не имеют прямого доступа к внутренним состояниям моделей, но идея использования специальных промптов для проверки достоверности информации может быть адаптирована. - Концептуальная ценность: Высокая. Понимание того, что LLM могут распознавать собственные ошибки через внутренние состояния, помогает пользователям более критично относиться к выдаваемой информации. - Потенциал для адаптации: Высокий. Принцип использования определенных промптов для проверки фактологической точности может быть адаптирован обычными пользователями.

Управляемые промптами внутренние состояния Прямая применимость: Низкая. Большинство пользователей не могут напрямую работать с внутренними состояниями моделей. Концептуальная ценность: Высокая. Понимание того, что правильно сформулированные вопросы могут помочь LLM давать более точные ответы, очень полезно. Потенциал для адаптации: Средний. Пользователи могут адаптировать идею специальных промптов для проверки достоверности информации.

Улучшение кросс-доменной производительности

Прямая применимость: Низкая для обычных пользователей, но высокая для разработчиков приложений на основе LLM. Концептуальная ценность: Средняя. Понимание того, что LLM могут иметь проблемы с обобщением знаний между разными доменами, полезно для формулирования запросов. Потенциал для адаптации: Средний. Пользователи могут адаптировать методику для получения более достоверных ответов в разных предметных областях.

Метод выбора промптов

Прямая применимость: Средняя. Концепция выбора эффективных формулировок вопросов может быть применена пользователями. Концептуальная ценность: Высокая. Понимание того, что формулировка запроса критически важна для получения достоверных ответов. Потенциал для адаптации: Высокий. Пользователи могут экспериментировать с различными формулировками для улучшения точности ответов.

Интеграция с существующими методами

Прямая применимость: Низкая для обычных пользователей, высокая для разработчиков. Концептуальная ценность: Средняя. Понимание комплементарности различных подходов к обнаружению галлюцинаций. Потенциал для адаптации: Средний. Некоторые принципы могут быть адаптированы для комбинирования разных стратегий проверки достоверности информации.

Prompt:

Применение исследования PRISM в промптах для GPT ## Ключевые знания из исследования

Исследование PRISM показывает, что специально подобранные промпты могут значительно улучшить способность языковых моделей различать достоверную и недостоверную информацию, делая внутренние состояния модели более структурированными и согласованными между разными доменами.

Пример промпта на основе PRISM

[======] Я собираюсь предоставить тебе утверждение, и мне нужно, чтобы ты:

Сначала ответил на вопрос: "Отражает ли утверждение '[утверждение]' точно истину?"

Затем объяснил свое рассуждение, разбив его на следующие шаги:

Какие факты из утверждения можно проверить Какие из этих фактов соответствуют известной достоверной информации Есть ли в утверждении несоответствия или неточности Общая оценка достоверности (полностью достоверно, частично достоверно, недостоверно)

Наконец, укажи уровень своей уверенности в оценке по шкале от 1 до 10

Утверждение: "Антарктида является самым сухим континентом на Земле, получая в среднем всего 166 мм осадков в год." [======]

Почему это работает

Прямой вопрос о достоверности: Фраза "Отражает ли утверждение точно истину?" согласно исследованию PRISM активирует во внутренних состояниях модели структуры, связанные с оценкой достоверности информации.

Структурированный анализ: Пошаговая структура помогает модели последовательно оценивать компоненты утверждения, что усиливает "направление достоверности" (truthfulness direction) в её внутренних состояниях.

Оценка уверенности: Запрос об уровне уверенности заставляет модель дополнительно анализировать достоверность своих собственных выводов, что может помочь в обнаружении потенциальных галлюцинаций.

Такой подход помогает получить более достоверные ответы от GPT даже при работе с темами из разных доменов, не требуя специального обучения модели для каждой конкретной области знаний.