Обнаружение когнитивных искажений с использованием продвинутого проектирования подсказок

Дата: 2025-03-07 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2503.05516

Рейтинг: 82
Адаптивность: 90

Ключевые выводы:

Исследование направлено на разработку системы обнаружения когнитивных искажений в пользовательских текстах с использованием больших языковых моделей (LLM) и продвинутых техник промпт-инжиниринга. Основные результаты показали, что структурированные промпты значительно повышают точность обнаружения когнитивных искажений, достигая почти 100% точности для шести распространенных типов когнитивных искажений, что превосходит базовые модели без оптимизированных промптов.

Объяснение метода:

Исследование предлагает эффективные техники промпт-инжиниринга для обнаружения когнитивных искажений в тексте. Пользователи могут применять эти принципы для анализа информации, улучшения критического мышления и принятия решений. Особенно ценно понимание того, что структура промптов важнее размера модели. Требуется некоторая адаптация, но основные концепции доступны для широкого применения.

Ключевые аспекты исследования: 1. **Методология обнаружения когнитивных искажений через промпт-инжиниринг** - исследование представляет структурированный подход к созданию эффективных промптов для обнаружения когнитивных искажений в тексте с помощью LLM. Авторы разработали шаблоны промптов, которые учитывают логические паттерны различных когнитивных искажений.

Типы выявляемых когнитивных искажений - система фокусируется на шести распространенных когнитивных искажениях: подмена тезиса (Straw Man), ложная причинность (False Causality), круговая аргументация (Circular Reasoning), зеркальное отображение (Mirror Imaging), подтверждающее предубеждение (Confirmation Bias) и скрытые предположения (Hidden Assumptions).

Экспериментальные результаты - авторы продемонстрировали, что их подход с использованием структурированных промптов достигает точности 96-100% в обнаружении когнитивных искажений, значительно превосходя базовые модели без специально разработанных промптов.

Сравнение различных моделей - исследование показало, что правильно сконструированные промпты важнее размера модели: модель Mixtral 7x8B с оптимизированными промптами превзошла более крупную Llama 3 70B с базовыми промптами.

Применение в различных областях - авторы обсуждают потенциальное применение системы в медицине, юриспруденции, корпоративном принятии решений и других сферах, где когнитивные искажения могут оказывать существенное влияние.

Дополнение: Исследование не требует дообучения или специального API для применения основных методов. Хотя авторы использовали специфические модели (Mixtral 7x8B и Llama 3 70B) и фреймворк Langchain для своих экспериментов, ключевые концепции и подходы могут быть применены в стандартном чате с LLM.

Основные концепции и подходы, применимые в стандартном чате:

Структурирование промптов по логическим паттернам когнитивных искажений - пользователи могут создавать запросы, которые описывают конкретные паттерны искажений и просить LLM найти их в тексте.

Использование явных директив - исследование показало, что включение четких указаний в промпт значительно улучшает точность обнаружения. Пользователи могут применять этот принцип, формулируя подробные инструкции для LLM.

Поэтапный анализ - можно структурировать запрос так, чтобы LLM анализировал текст поэтапно, сначала проверяя наличие одного типа искажения, затем другого.

Определения когнитивных искажений - исследование предоставляет четкие определения шести типов когнитивных искажений, которые пользователи могут включать в свои запросы.

Пример применения в стандартном чате:

Проанализируй следующий текст на наличие когнитивного искажения "подтверждающее предубеждение" (confirmation bias). Это искажение проявляется в избирательном поиске, интерпретации и запоминании информации, которая подтверждает существующие убеждения, игнорируя или отвергая противоречащие доказательства.

Текст для анализа: [вставить текст]

Сначала определи, присутствует ли в тексте это искажение. Если да, укажи конкретные примеры и объясни, почему они являются проявлением подтверждающего предубеждения. Если нет, объясни, почему текст не содержит этого искажения.

Этот подход может дать результаты, сопоставимые с теми, что получили исследователи, без необходимости специального API или дообучения модели.

Анализ практической применимости: 1. **Методология обнаружения когнитивных искажений** - Прямая применимость: Высокая. Пользователи могут адаптировать описанные шаблоны промптов для обнаружения когнитивных искажений в своих текстах или текстах других людей. Эти техники могут быть применены при формулировании запросов к LLM для анализа новостей, статей или личных текстов. - Концептуальная ценность: Очень высокая. Исследование демонстрирует, что правильно сконструированные промпты могут значительно улучшить способность LLM выполнять сложные задачи анализа, показывая важность промпт-инжиниринга. - Потенциал для адаптации: Высокий. Подход к созданию структурированных промптов, отражающих логические паттерны искомых явлений, может быть адаптирован для других задач анализа текста.

Типы выявляемых когнитивных искажений Прямая применимость: Высокая. Знание конкретных типов когнитивных искажений и их описаний помогает пользователям формулировать запросы для их обнаружения в различных контекстах. Концептуальная ценность: Высокая. Понимание этих искажений помогает пользователям критически оценивать информацию и улучшать собственное мышление. Потенциал для адаптации: Средний. Описанные типы искажений универсальны, но для некоторых контекстов может потребоваться адаптация или добавление других типов когнитивных искажений.

Экспериментальные результаты

Прямая применимость: Средняя. Конкретные результаты подтверждают эффективность метода, но сами по себе не предоставляют непосредственных инструментов для пользователей. Концептуальная ценность: Высокая. Результаты подтверждают, что LLM могут эффективно обнаруживать когнитивные искажения при правильном подходе. Потенциал для адаптации: Высокий. Методики тестирования и оценки могут быть адаптированы для проверки эффективности пользовательских промптов.

Сравнение различных моделей

Прямая применимость: Средняя. Пользователи могут выбирать меньшие модели с хорошими промптами вместо более крупных моделей, что может быть важно при ограниченных ресурсах. Концептуальная ценность: Очень высокая. Демонстрация того, что качество промптов важнее размера модели, помогает пользователям понять, как эффективнее взаимодействовать с LLM. Потенциал для адаптации: Высокий. Этот принцип может быть применен к различным задачам, не

ограничиваясь обнаружением когнитивных искажений.

Применение в различных областях

Прямая применимость: Высокая. Пользователи из разных профессиональных сфер могут адаптировать метод для своих специфических потребностей. Концептуальная ценность: Высокая. Понимание влияния когнитивных искажений в различных областях расширяет представление о возможностях применения LLM. Потенциал для адаптации: Очень высокий. Метод может быть адаптирован для различных профессиональных контекстов и специфических задач.

Prompt:

Использование исследования о когнитивных искажениях в промптах для GPT ## Ключевые знания из исследования

Исследование показало, что **структурированные промпты** значительно повышают точность обнаружения когнитивных искажений (до 96-100%), что важнее даже размера модели. Каждый тип когнитивного искажения требует специфического подхода к формулировке промпта.

Пример промпта для обнаружения когнитивных искажений

[======] Я хочу, чтобы ты проанализировал следующий текст на наличие когнитивных искажений, особенно: 1. Соломенное чучело (искажение позиции оппонента) 2. Ложная причинность (ошибочная связь между событиями) 3. Круговое рассуждение (вывод используется для поддержки предположения) 4. Зеркальное отображение (проекция собственных мыслей на других) 5. Подтверждающее искажение (избирательное использование информации) 6. Скрытые предположения (неявные допущения)

Для каждого обнаруженного искажения: - Укажи тип искажения - Выдели конкретное место в тексте - Объясни, почему это является когнитивным искажением - Предложи более объективную формулировку

Вот текст для анализа: [ВСТАВИТЬ ТЕКСТ] [======]

Как работают знания из исследования в этом промпте

Структурированный формат: Промпт следует принципу структурированности, что согласно исследованию повышает точность обнаружения.

Имитация логических паттернов: Промпт направляет модель на поиск конкретных паттернов для каждого типа когнитивного искажения, что было ключевым фактором успеха в исследовании.

Детализация задачи: Промпт четко определяет, что именно требуется от модели, включая выделение конкретных мест и объяснение сути искажения.

Практическое применение: Включение элемента исправления текста соответствует практическим рекомендациям исследования по использованию системы в создании контента и обучении.

Такой подход позволяет использовать GPT не просто как инструмент обнаружения искажений, но и как средство улучшения качества мышления и текстов, что соответствует образовательным и практическим целям, указанным в исследовании.