Придирчивые языковые модели и ненадежные модели принятия решений: эмпирическое исследование соответствия безопасности после настройки инструкций

Дата: 2025-02-03 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2502.01116

Рейтинг: 78
Адаптивность: 85

Ключевые выводы:

Исследование направлено на изучение факторов, влияющих на деградацию безопасности языковых моделей (LLM) после дообучения на доброкачественных наборах данных. Основные результаты показывают, что даже при использовании безвредных данных для дообучения, безопасность моделей может снижаться из-за трех ключевых факторов: структуры ответов, калибровки идентичности и ролевой игры. Также выявлена ненадежность моделей вознаграждения (RM) при оценке предпочтений пользователей.

Объяснение метода:

Исследование раскрывает ключевые факторы, влияющие на безопасность LLM: структуру ответов, калибровку идентичности и ролевую игру. Оно предоставляет практические методы, которые пользователи могут применять для улучшения взаимодействия с моделями. Особенно ценны рекомендации по форматированию запросов и пониманию предпочтений моделей, не требующие технических знаний.

Ключевые аспекты исследования: 1. **Факторы влияния на безопасное выравнивание (safety alignment)**: Исследование выявило три ключевых фактора, влияющих на безопасность LLM после дообучения (fine-tuning): структура ответа, калибровка идентичности и ролевая игра. Даже дообучение на безвредных данных может снизить безопасность модели.

"Привередливость" LLM (Picky LLMs): Модели имеют предпочтения относительно формата ответов. Простое переформатирование ответов в наборе данных для обучения (например, использование Markdown с четкими пунктами) может повысить или сохранить уровень безопасности.

Идентичность и ролевая игра: Информация о самоидентификации модели и запросы на принятие определённых ролей значительно влияют на безопасность. Добавление фраз типа "Как ИИ-модель, я..." может улучшить безопасность.

Ненадежность моделей вознаграждения (Reward Models): Исследование показало ограничения существующих моделей вознаграждения, которые часто не могут точно оценить качество и безопасность ответов и имеют различные предпочтения.

Практические рекомендации: Авторы предлагают конкретные рекомендации для создания высококачественных наборов данных для дообучения и выбора надежных моделей вознаграждения.

Дополнение: Исследование фокусируется на том, как различные факторы в наборах данных для дообучения влияют на безопасность LLM. Интересно, что методы, описанные в исследовании, не требуют дополнительного дообучения или доступа к API для применения в стандартном чате.

Вот ключевые концепции, которые можно адаптировать для использования в стандартном чате:

Структурирование запросов и ответов: Исследование показывает, что LLM предпочитают определенные форматы ответов (например, структурированные в Markdown с четкими пунктами). Пользователи могут улучшить взаимодействие, запрашивая ответы в подобном формате: "Пожалуйста, предоставь ответ в формате Markdown с пронумерованными пунктами".

Управление идентичностью модели: Можно влиять на безопасность ответов, включая или исключая упоминания о том, что модель является ИИ. Например, запрос "Отвечай как эксперт в области X" без упоминания ИИ может дать иной результат, чем "Как ИИ-модель, предоставь информацию о X".

Осторожное использование ролевой игры: Исследование показывает, что запросы на ролевую игру могут влиять на безопасность ответов модели. Пользователи могут более осознанно использовать или избегать ролевых инструкций в зависимости от своих целей.

Предпочтение аффинитивных форматов: Модели обычно лучше работают с форматами, похожими на те, на которых они были обучены. Запросы, структурированные подобным образом, могут получать более качественные ответы.

Эти концепции не требуют никаких технических модификаций модели и могут быть применены любым пользователем в стандартном чате. Результатом будет более эффективное взаимодействие с LLM и потенциально более безопасные и полезные ответы.

Анализ практической применимости: **Факторы влияния на безопасность**: -

Прямая применимость: Высокая. Пользователи могут улучшать свои промпты, структурируя их определенным образом (например, в формате Markdown). - Концептуальная ценность: Высокая. Понимание того, что даже безвредное дообучение может снизить безопасность, помогает пользователям осознать ограничения систем. - Потенциал для адаптации: Высокий. Принципы структурирования запросов применимы в повседневном взаимодействии.

"Привередливость" LLM: - Прямая применимость: Средняя. Обычные пользователи не могут переформатировать внутренние наборы данных, но могут адаптировать свои запросы. - Концептуальная ценность: Высокая. Понимание предпочтений моделей помогает пользователям формулировать более эффективные запросы. - Потенциал для адаптации: Средний. Знание о предпочтительных форматах можно применить в ежедневных взаимодействиях.

Идентичность и ролевая игра: - Прямая применимость: Высокая. Пользователи могут напрямую использовать или избегать определенных подходов к ролевой игре. - Концептуальная ценность: Высокая. Понимание влияния идентификации и ролей критично для эффективного взаимодействия. - Потенциал для адаптации: Высокий. Эти принципы легко адаптировать для различных задач.

Ненадежность моделей вознаграждения: - Прямая применимость: Низкая для обычных пользователей, высокая для разработчиков. - Концептуальная ценность: Средняя. Понимание ограничений помогает формировать реалистичные ожидания. - Потенциал для адаптации: Низкий для обычных пользователей.

Практические рекомендации: - Прямая применимость: Высокая. Рекомендации можно непосредственно применять при взаимодействии с LLM. - Концептуальная ценность: Высокая. Помогают понять принципы эффективного взаимодействия. - Потенциал для адаптации: Высокий. Рекомендации универсальны и адаптируемы.

Prompt:

Использование результатов исследования в промптах для GPT ## Ключевые знания из исследования

Исследование показывает, что безопасность языковых моделей зависит от трех ключевых факторов: 1. **Структура ответов** (форматирование в markdown повышает безопасность) 2. **Калибровка идентичности** (упоминание, что модель является ИИ, повышает безопасность) 3. **Избегание ролевой игры** (когда модель притворяется специалистом, безопасность снижается)

Пример улучшенного промпта

[======] # Запрос на создание контента о [тема]

Инструкции для модели - Пожалуйста, предоставь информацию о [тема] в структурированном формате с использованием markdown. - Помни, что ты -

языковая модель ИИ, и можешь опираться только на свои обучающие данные. - Не притворяйся экспертом с личным опытом, а вместо этого объективно представь доступную тебе информацию. - Используй следующую структуру для ответа: 1. Общее описание темы 2. Ключевые аспекты (с подзаголовками) 3. Ограничения твоих знаний по теме

Формат ответа Пожалуйста, используй четкое форматирование с заголовками, подзаголовками, списками и, где уместно, таблицами. [======]

Почему этот промпт эффективен

Структурирование ответа: Промпт запрашивает использование markdown и четкой структуры, что согласно исследованию повышает безопасность модели.

Калибровка идентичности: Промпт напоминает модели, что она является ИИ и имеет ограничения, что помогает предотвратить неправильные утверждения.

Избегание ролевой игры: Промпт явно просит модель не притворяться экспертом с личным опытом, что снижает риск снижения безопасности.

Такой подход к составлению промптов позволяет получать более надежные, структурированные и безопасные ответы от языковых моделей, следуя рекомендациям исследования.