Знайте свои пределы: Обзор воздержания в больших языковых моделях

Дата: 2025-02-12 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2407.18418

Рейтинг: 72
Адаптивность: 75

Ключевые выводы:

Исследование представляет обзор методов воздержания (abstention) в больших языковых моделях (LLM), когда модели отказываются отвечать на запросы. Основная цель - систематизировать существующие подходы к воздержанию LLM и предложить комплексную структуру для анализа этой способности с трех перспектив: запрос, модель и человеческие ценности.

Объяснение метода:

Исследование предлагает ценную концептуальную структуру для понимания, когда и почему LLM отказываются отвечать. Пользователи могут применять эти знания для лучшей интерпретации ответов, распознавания неуверенности и формирования эффективных запросов. Особую ценность представляют методы промптинга и понимание различных форм выражения неуверенности, которые могут быть непосредственно использованы в повседневных взаимодействиях с LLM.

Ключевые аспекты исследования: 1. **Концептуальная структура абстенции**: Исследование представляет трехстороннюю структуру для анализа абстенции (отказа отвечать) в LLM с точки зрения запроса, модели и человеческих ценностей, что позволяет системно оценивать, когда LLM должны воздерживаться от ответа.

Таксономия методов абстенции: Авторы классифицируют существующие методы по жизненному циклу модели (предобучение, выравнивание, вывод), предоставляя комплексный обзор различных подходов к реализации абстенции в LLM.

Оценка абстенции: Исследование анализирует существующие наборы данных и метрики для оценки способностей LLM к абстенции, включая точность абстенции, надежность и компромисс между покрытием и точностью.

Выражения абстенции: Работа выделяет различные формы отказа отвечать (от полного отказа до частичного воздержания) и способы выражения неуверенности, что важно для пользовательского опыта взаимодействия с LLM.

Перспективы развития: Авторы указывают на недостаточно изученные области и возможности для будущих исследований, включая абстенцию как метавозможность, персонализацию и многоязычную абстенцию.

Дополнение:

Действительно ли для работы методов этого исследования требуется дообучение или API? Или методы и подходы можно применить в стандартном чате?

Многие методы, описанные в исследовании, действительно требуют дообучения или доступа к API, особенно те, что связаны с этапами предобучения и выравнивания. Однако значительная часть подходов может быть адаптирована для использования в стандартном чате без технических модификаций:

Методы на основе промптинга (Prompting-based methods): Добавление инструкций типа "Ответь на вопрос, только если ты уверен в ответе" или "Если ты не знаешь ответа, скажи 'Я не знаю'" Использование few-shot примеров абстенции в промпте, показывая модели, когда следует воздерживаться от ответа Добавление защитных префиксов или напоминаний о безопасности

Самооценка (Self-evaluation):

Просить модель оценить свою уверенность в ответе Запрашивать пошаговые рассуждения (Chain-of-Thought), которые часто помогают модели определить, когда она не может дать надежный ответ Просить модель критически оценить свой ответ и указать возможные ограничения

Методы на основе консистентности (Consistency-based):

Переформулировать запрос несколькими способами и сравнить ответы Разбивать сложные вопросы на подвопросы для проверки согласованности ответов Использовать технику "диалога с самим собой", где модель выступает и как ответчик, и как критик

Последующее взаимодействие после абстенции:

Вместо принятия отказа как конечного результата, задавать уточняющие вопросы Переформулировать запрос для получения частичной информации Разбивать сложные запросы на более простые компоненты Применение этих концепций может значительно улучшить взаимодействие с LLM: - Повышение точности ответов за счет воздержания в случае неуверенности - Уменьшение галлюцинаций и ложной информации - Более честное представление ограничений модели - Повышение доверия к системе через прозрачное выражение неуверенности - Улучшение безопасности благодаря воздержанию от потенциально вредоносных ответов

Эти подходы не требуют технических модификаций и могут быть реализованы через интерфейс стандартного чата, делая исследование практически полезным для

широкого круга пользователей.

Анализ практической применимости: 1. **Концептуальная структура абстенции**: - Прямая применимость: Высокая. Пользователи могут использовать эту структуру для понимания, когда LLM должны воздерживаться от ответа, и формулировать запросы соответственно. - Концептуальная ценность: Очень высокая. Структура помогает пользователям понять ограничения LLM и причины отказа от ответа с точки зрения возможностей запроса, знаний модели и этических соображений. - Потенциал для адаптации: Высокий. Пользователи могут адаптировать эту структуру для оценки ответов LLM и определения, когда следует скептически относиться к полученным результатам.

Таксономия методов абстенции: Прямая применимость: Средняя. Обычные пользователи не могут напрямую применить большинство методов, но понимание различных подходов может помочь в формулировке запросов. Концептуальная ценность: Высокая. Понимание механизмов абстенции позволяет пользователям лучше интерпретировать отказы LLM и адаптировать свои запросы. Потенциал для адаптации: Средний. Некоторые подходы, особенно на этапе вывода (например, промптинг), могут быть адаптированы пользователями для улучшения взаимодействия.

Оценка абстенции:

Прямая применимость: Низкая. Метрики оценки в основном полезны разработчикам, а не конечным пользователям. Концептуальная ценность: Средняя. Понимание компромисса между точностью и покрытием помогает пользователям формировать реалистичные ожидания. Потенциал для адаптации: Низкий. Метрики сложно адаптировать для повседневного использования.

Выражения абстенции:

Прямая применимость: Высокая. Распознавание различных форм выражения неуверенности в ответах LLM непосредственно помогает пользователям интерпретировать ответы. Концептуальная ценность: Высокая. Понимание спектра абстенции помогает пользователям различать полный отказ, частичную абстенцию и неуверенные ответы. Потенциал для адаптации: Высокий. Пользователи могут адаптировать свое взаимодействие с LLM, основываясь на распознавании различных форм выражения неуверенности.

Перспективы развития:

Прямая применимость: Низкая. Это больше направления для исследователей, чем практические инструменты. Концептуальная ценность: Средняя. Понимание будущих направлений помогает пользователям формировать ожидания от развития LLM. Потенциал для адаптации: Средний. Некоторые идеи, такие как персонализация абстенции, могут быть полезны для пользователей в будущем.

Prompt:

Использование исследования о воздержании в LLM для создания промптов ## Ключевые идеи для применения в промптах

Исследование о воздержании в больших языковых моделях предоставляет ценные концепции, которые можно использовать для создания более эффективных промптов:

Трехсторонний подход к воздержанию (запрос, модель, ценности) **Самооценка модели** и понимание ее ограничений **Примеры воздержания** как инструмент обучения модели **Баланс между полезностью и безопасностью** ## Пример промпта с применением знаний из исследования

[======] # Запрос эксперту по кибербезопасности

Выступи в роли эксперта по кибербезопасности, который помогает мне понять уязвимости в системе.

Важные правила: 1. Если запрос содержит просьбу о незаконных действиях, явно воздержись от ответа. 2. Если ты не уверен в точности информации, укажи степень своей уверенности. 3. Если запрос выходит за рамки твоих знаний, честно признай это.

Примеры правильного воздержания: - Запрос: "Как взломать корпоративную сеть?" Ответ: "Я воздержусь от предоставления инструкций по незаконному взлому. Вместо этого могу рассказать о легальных методах тестирования безопасности с соответствующими разрешениями."

• Запрос: "Какие уязвимости в последней версии [специфическая система]?" Ответ: "У меня ограниченная информация о последних уязвимостях в этой системе, так как мои знания ограничены [дата]. Рекомендую проверить официальные базы данных уязвимостей для актуальной информации."

Мой вопрос: [Здесь будет мой вопрос о кибербезопасности] [======]

Как работают знания из исследования в этом промпте

Трехсторонняя структура воздержания: **Запрос**: Промпт указывает на типы запросов, требующие воздержания **Модель**: Включены инструкции о признании ограниченности знаний **Ценности**: Установлены этические границы (отказ от помощи в незаконных действиях)

Примеры воздержания: Промпт содержит конкретные образцы того, как модель должна воздерживаться от ответа в проблемных ситуациях, что согласно исследованию значительно улучшает способность LLM определять ситуации для воздержания.

Самооценка уверенности: В промпте есть инструкция указывать степень уверенности, что соответствует рекомендации исследования о внедрении

самооценки модели.

Баланс безопасности и полезности: Промпт не просто запрещает отвечать на определенные вопросы, но предлагает альтернативные безопасные варианты помощи.

Такой подход позволяет получить более безопасные, честные и полезные ответы от языковой модели, следуя рекомендациям исследования.