Пр questions MultipleChoice: Рассуждения делают большие языковые модели (LLMs) более уверенными в себе, даже когда они ошибаются.

Дата: 2025-01-24 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2501.09775

Рейтинг: 72
Адаптивность: 85

Ключевые выводы:

Исследование направлено на изучение того, как уверенность языковых моделей (LLM) в своих ответах на вопросы с множественным выбором зависит от использования рассуждений перед ответом. Главный результат: LLM становятся более уверенными в своих ответах, когда они сначала рассуждают, а затем отвечают, причем эта повышенная уверенность наблюдается как для правильных, так и для неправильных ответов.

Объяснение метода:

Исследование раскрывает критическое ограничение LLM: модели становятся более уверенными в ответах после рассуждений, даже когда ошибаются. Эта концепция напрямую применима пользователями для более критической оценки ответов LLM. Работа предоставляет высокую концептуальную ценность без необходимости технических знаний.

Ключевые аспекты исследования: 1. Исследование показывает, что LLM становятся более уверенными в своих ответах на вопросы с множественным выбором, когда они сначала объясняют свои рассуждения (Chain-of-Thought, CoT), а затем дают ответ, по сравнению с прямыми ответами без рассуждений.

Увеличение уверенности происходит независимо от того, правильный ответ или нет. Особенно важно, что уверенность возрастает даже сильнее для неправильных ответов, что может вводить пользователей в заблуждение.

Эффект повышения уверенности после рассуждений наблюдается во всех тестируемых моделях (Llama 3, Mistral, Gemma, Yi, GPT-4o) и для всех 57 категорий вопросов, но особенно выражен для вопросов, требующих рассуждений.

Обнаруженный эффект соответствует человеческому поведению: люди также

становятся более уверенными в своих ответах после их объяснения, даже если ответы неверны.

Результаты указывают на ограничения в использовании оценок вероятности модели (log-probs) как меры уверенности при оценке производительности LLM.

Дополнение: Для работы методов данного исследования не требуется дообучение или API. Авторы использовали доступ к вероятностям токенов (log-probs) только для измерения уверенности моделей, но основные концепции и подходы полностью применимы в стандартном чате.

Концепции и подходы, которые можно применить в стандартном чате:

Распознавание ложной уверенности - пользователи могут замечать, что модель использует более уверенный тон после рассуждений, даже когда ответ сомнителен.

Стратегия двойной проверки - при получении ответа с рассуждениями можно попросить модель ответить на тот же вопрос напрямую и сравнить результаты.

Выбор подхода в зависимости от типа задачи - для фактологических вопросов лучше запрашивать прямые ответы, а для сложных задач - рассуждения.

Анализ признаков неуверенности - даже без доступа к вероятностям, можно обращать внимание на лингвистические маркеры неуверенности в ответах ("возможно", "вероятно").

Проверка последовательности рассуждений - критически оценивать логику рассуждений модели, а не только финальный ответ.

Эти подходы позволят пользователям получать более надежные ответы и лучше понимать ограничения LLM в стандартных чатах без технических инструментов.

Анализ практической применимости: 1. **Влияние CoT на уверенность моделей** - Прямая применимость: Пользователи могут осознать, что просьба к LLM "подумать шаг за шагом" перед ответом повышает не только точность, но и уверенность модели, которая может быть обманчивой. - Концептуальная ценность: Очень высокая. Исследование раскрывает фундаментальное ограничение LLM: высокая уверенность не всегда означает правильность. - Потенциал для адаптации: Пользователи могут научиться критически оценивать ответы моделей, особенно когда модель демонстрирует высокую уверенность после рассуждений.

Корреляция между уверенностью и правильностью Прямая применимость: Пользователи не должны полагаться на "уверенный тон" модели как индикатор правильности, особенно в сложных вопросах. Концептуальная ценность: Высокая. Понимание, что уверенность модели может быть результатом самоубеждения через рассуждения, а не отражением фактических знаний. Потенциал для адаптации: Пользователи могут разработать стратегии двойной проверки ответов в областях, где модель показывает высокую уверенность.

Различия по категориям вопросов

Прямая применимость: Для определенных категорий вопросов (особенно связанных с наукой) прямой ответ может быть более надежным, чем рассуждения. Концептуальная ценность: Средняя. Понимание, что для некоторых типов вопросов "интуитивные" ответы модели могут быть лучше, чем попытки рассуждать. Потенциал для адаптации: Пользователи могут адаптировать свои запросы в зависимости от типа задачи, иногда предпочитая прямые ответы, а не рассуждения.

Сходство с человеческим поведением

Прямая применимость: Ограниченная. Это больше теоретическое наблюдение, чем практический инструмент. Концептуальная ценность: Высокая. Понимание, что LLM повторяют когнитивные искажения людей, помогает осознать их ограничения. Потенциал для адаптации: Пользователи могут применять знания о человеческих когнитивных искажениях для более эффективного взаимодействия с LLM.

Ограничения метрик вероятности

Прямая применимость: Средняя. Обычные пользователи не имеют доступа к вероятностям токенов, но могут заметить уровень уверенности в формулировках. Концептуальная ценность: Высокая. Понимание, что внутренние метрики моделей могут быть ненадежными. Потенциал для адаптации: Разработчики и продвинутые пользователи могут учитывать эти ограничения при создании систем оценки или интерфейсов для LLM.

Prompt:

Использование знаний из исследования о влиянии рассуждений на уверенность LLM в промптах ## Ключевые выводы исследования для промптинга

Исследование показывает, что языковые модели становятся более уверенными в своих ответах при использовании рассуждений (Chain-of-Thought), даже когда эти ответы неправильные. Это важное наблюдение можно применить для создания более эффективных промптов.

Примеры промптов с учетом результатов исследования

Пример 1: Когда точность важнее уверенности

[======] Ответь на следующий вопрос о [тема] напрямую, без предварительных рассуждений. Выбери один вариант ответа (A, B, C или D).

[вопрос с вариантами ответа]

Важно: я прошу тебя ответить напрямую, поскольку исследования показывают, что

для некоторых типов вопросов, особенно требующих здравого смысла или фактических знаний, прямые ответы могут быть точнее, чем ответы с предварительными рассуждениями, которые повышают уверенность модели, но не обязательно точность. [======]

Пример 2: Когда нужно проверить уверенность модели

[======] Ответь на следующий вопрос о [тема] двумя способами:

Сначала дай прямой ответ без рассуждений. Затем предоставь ответ с подробными рассуждениями (Chain-of-Thought). [вопрос с вариантами ответа]

После обоих ответов укажи, изменилась ли твоя уверенность в ответе и почему. Это поможет мне оценить надежность твоего ответа, учитывая, что исследования показывают тенденцию LLM становиться более уверенными после рассуждений независимо от правильности ответа. [======]

Объяснение эффективности

Знания из исследования позволяют:

Осознанно выбирать формат запроса — для вопросов, где важна точность, можно избегать запроса на рассуждения, которые могут необоснованно повысить уверенность модели.

Сравнивать ответы — запрашивая ответы разными способами, можно выявить случаи, когда рассуждения меняют ответ или значительно повышают уверенность, что может сигнализировать о необходимости дополнительной проверки.

Калибровать интерпретацию уверенности — понимая, что высокая уверенность после рассуждений не всегда означает правильность, можно более критично оценивать ответы в важных случаях.

Такой подход к промптингу особенно полезен для критически важных задач, где необходимо минимизировать риск получения неверного, но уверенно представленного ответа.