Когда AI беспокоится о своих ответах — и когда его неопределенность оправдана

Дата: 2025-03-03 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2503.01688

Рейтинг: 68
Адаптивность: 75

Ключевые выводы:

Исследование направлено на оценку методов определения неопределенности (uncertainty estimation) в ответах больших языковых моделей (LLM) при решении задач с множественным выбором. Основной вывод: энтропия токенов хорошо предсказывает ошибки модели в задачах, требующих знаний (ROC AUC 0.73 для биологии), но плохо работает для задач, требующих рассуждений (ROC AUC 0.55 для математики).

Объяснение метода:

Исследование предоставляет практический метод (энтропия) для оценки достоверности ответов LLM, особенно в задачах, требующих знаний. Результаты помогают пользователям понять, в каких типах задач LLM более надежны, и критически оценивать высокую заявленную уверенность. Однако применение требует технических знаний, а исследование ограничено вопросами с множественным выбором.

Ключевые аспекты исследования: 1. **Исследование оценки неопределенности LLM**: Авторы изучают, как различные методы оценки неопределенности (энтропия токенов и Model-as-Judge) работают для задач с вопросами с множественным выбором по разным темам. 2. **Корреляция энтропии с ошибками модели**: Установлено, что энтропия ответа хорошо предсказывает ошибки модели в областях, зависящих от знаний (биология, ROC AUC 0.73), но эта корреляция исчезает для задач, требующих рассуждений (математика, ROC AUC 0.55). 3. **Зависимость от размера модели**: Более крупные модели (Qwen-72B) демонстрируют лучшую способность оценивать собственную неопределенность через энтропию (ROC AUC 0.77), чем меньшие модели. 4. **Влияние типа задачи**: Энтропия лучше предсказывает ошибки в вопросах, требующих знаний, а не рассуждений, что указывает на разные типы неопределенности в разных задачах. 5. **Проблемы калибровки**: Все модели демонстрируют систематическую переоценку своей уверенности, особенно в областях с высокой заявленной уверенностью.

Дополнение:

Применимость методов исследования в стандартном чате

Исследование не требует дообучения модели или специального API для применения основных концепций. Хотя авторы использовали прямой доступ к логитам модели для расчета энтропии, концептуальные выводы можно адаптировать для стандартного чата:

Оценка уверенности через перефразирование вопроса: Пользователи могут попросить модель ответить на один и тот же вопрос несколькими способами Если ответы согласованы, это указывает на низкую энтропию (высокую уверенность) Если ответы различаются, это указывает на высокую энтропию (низкую уверенность)

Явный запрос об уверенности:

Пользователи могут напрямую спросить: "Насколько ты уверен в этом ответе?" Можно попросить модель оценить свою уверенность по шкале от 1 до 10 Исследование показывает, что такие оценки следует интерпретировать с осторожностью, особенно в задачах, требующих рассуждений

Разделение сложных задач на задачи знаний и рассуждений:

Ключевой вывод исследования - LLM лучше оценивают свою уверенность в задачах на знания, чем в задачах на рассуждения Пользователи могут разбивать сложные вопросы на части: "Какие факты нам известны?" и "Какие выводы мы можем сделать из этих фактов?" Это позволяет отделить компоненты знаний (где оценка уверенности более надежна) от компонентов рассуждений

Применение MASJ через самопроверку:

Хотя MASJ показал слабые результаты в исследовании, концепцию можно адаптировать Пользователи могут попросить модель сначала ответить на вопрос, а затем критически оценить свой ответ Например: "Пожалуйста, ответь на этот вопрос, а затем объясни, какие аспекты твоего ответа могут быть неточными или требуют дополнительной проверки" Ожидаемые результаты от применения этих методов: - Более критическая оценка ответов модели пользователями - Снижение риска принятия неверных ответов с высокой заявленной уверенностью - Повышение осведомленности о различиях между задачами на знания и задачами на рассуждения - Более эффективное взаимодействие с LLM через формулирование запросов, учитывающих особенности оценки неопределенности

Анализ практической применимости: **1. Использование энтропии для оценки достоверности ответов** - **Прямая применимость**: Высокая. Пользователи могут запрашивать у LLM уровень уверенности в ответе или энтропию и использовать это как индикатор достоверности, особенно в задачах, требующих знаний, а не рассуждений. - **Концептуальная ценность**: Значительная. Понимание того, что

низкая энтропия коррелирует с правильными ответами, помогает пользователям оценивать надежность получаемой информации. - **Потенциал для адаптации**: Высокий. Пользователи могут запрашивать LLM оценить свою уверенность в разных частях ответа, что особенно полезно для сложных запросов.

2. Различия в оценке неопределенности для разных типов вопросов - **Прямая применимость**: Средняя. Пользователи могут учитывать, что LLM более надежны в вопросах, требующих знаний, чем в вопросах, требующих рассуждений. - **Концептуальная ценность**: Высокая. Понимание, что разные типы задач вызывают разные типы неопределенности, помогает пользователям формировать запросы соответствующим образом. - **Потенциал для адаптации**: Средний. Пользователи могут разбивать сложные рассуждения на более простые шаги, чтобы повысить надежность ответов.

3. Влияние размера модели на оценку неопределенности - **Прямая применимость**: Низкая для обычного пользователя, который не выбирает модель напрямую. - **Концептуальная ценность**: Средняя. Понимание того, что более крупные модели обычно лучше оценивают свою неуверенность. - **Потенциал для адаптации**: Низкий. Большинство пользователей не могут выбирать размер модели.

4. Проблемы калибровки и переоценка уверенности - **Прямая применимость**: Высокая. Пользователи должны относиться скептически к высокой уверенности модели, особенно в сложных вопросах. - **Концептуальная ценность**: Высокая. Понимание, что LLM склонны переоценивать свою уверенность, помогает пользователям критически оценивать ответы. - **Потенциал для адаптации**: Средний. Пользователи могут запрашивать альтернативные точки зрения или противоположные аргументы для проверки надежности ответов.

5. Использование MASJ для оценки сложности вопросов - **Прямая применимость**: Низкая. Метод показал слабые результаты в предсказании ошибок. - **Концептуальная ценность**: Средняя. Понимание того, что самооценка LLM не всегда надежна. - **Потенциал для адаптации**: Низкий. Требуется значительная доработка метода.

Сводная оценка полезности: Предварительная оценка: 65 баллов

Исследование предоставляет ценные концепции и методы для оценки неопределенности в ответах LLM, которые могут быть непосредственно применены пользователями. Особенно полезен вывод о том, что энтропия хорошо коррелирует с правильностью ответов в задачах, требующих знаний, но не в задачах, требующих рассуждений. Это дает пользователям практический инструмент для оценки надежности ответов.

Контраргументы к оценке: 1. **Почему оценка могла быть выше**: Исследование предоставляет конкретный метод (энтропия), который может быть адаптирован для повседневного использования и помогает пользователям понять, когда доверять LLM. Методология исследования также может быть использована для проверки надежности других моделей.

Почему оценка могла быть ниже: Исследование фокусируется на вопросах с множественным выбором, что ограничивает его применимость к более общим случаям использования LLM. Метод MASJ показал низкую эффективность, а использование энтропии требует технических знаний и не всегда доступно в стандартных интерфейсах LLM. После рассмотрения контраргументов, корректирую оценку до 68 баллов, признавая высокую ценность основных выводов, но учитывая некоторые ограничения в их непосредственном применении рядовыми пользователями.

Оценка в 68 баллов обоснована следующими факторами: 1. Исследование предоставляет практический метод (энтропия) для оценки достоверности ответов LLM. 2. Результаты помогают пользователям понять, в каких типах задач LLM более надежны. 3. Выводы о влиянии размера модели и типа задачи имеют практическую ценность. 4. Однако применение энтропии требует технических знаний и не всегда доступно напрямую. 5. Исследование ограничено вопросами с множественным выбором, что снижает его общую применимость.

Уверенность в оценке: Очень сильная. Исследование предоставляет четкие количественные результаты, которые напрямую связаны с практическими сценариями использования LLM. Выводы логически следуют из данных и согласуются с существующими знаниями о работе LLM. Методология исследования хорошо описана и воспроизводима.

Оценка адаптивности: Оценка адаптивности: 75 из 100.

1) **Адаптация принципов**: Концепция использования энтропии ответа как показателя неопределенности может быть адаптирована для стандартного чата путем запроса модели оценить свою уверенность или представить альтернативные ответы. Хотя прямой доступ к энтропии обычно недоступен, можно использовать прокси-показатели, такие как разнообразие возможных ответов.

2) **Извлечение полезных идей**: Пользователи могут применять ключевое понимание, что LLM более надежны в задачах, требующих знаний, чем в задачах, требующих сложных рассуждений. Это может помочь им формулировать запросы и интерпретировать ответы соответствующим образом.

3) **Потенциал для внедрения**: Высокий потенциал для включения оценки неопределенности в интерфейсы LLM, например, путем предоставления пользователям индикаторов уверенности модели или выделения частей ответа с высокой/низкой уверенностью.

4) **Абстрагирование методов**: Принцип "запрашивать модель оценить свою уверенность" может быть применен к различным типам взаимодействий, не ограничиваясь вопросами с множественным выбором.

|| <Оценка: 68> || <Объяснение: Исследование предоставляет практический метод

(энтропия) для оценки достоверности ответов LLM, особенно в задачах, требующих знаний. Результаты помогают пользователям понять, в каких типах задач LLM более надежны, и критически оценивать высокую заявленную уверенность. Однако применение требует технических знаний, а исследование ограничено вопросами с множественным выбором.> || <Адаптивность: 75>

Prompt:

Использование исследования о неопределенности AI в промптах для GPT
Ключевые знания из исследования

Исследование показывает, что: - Языковые модели лучше определяют свою неуверенность в фактологических задачах, чем в задачах рассуждения - Энтропия токенов хорошо предсказывает ошибки в областях знаний (биология, психология) - Для задач рассуждения (математика, физика) стандартные методы определения неопределенности работают плохо

Пример промпта с использованием этих знаний

[======] Я хочу, чтобы ты решил следующую математическую задачу. Поскольку исследования показывают, что языковые модели могут испытывать трудности с оценкой собственной уверенности в задачах рассуждения, пожалуйста:

Раздели решение на четкие логические шаги После каждого шага укажи уровень уверенности (высокий/средний/низкий) Если возможно, предложи альтернативный подход к решению В конце оцени общую уверенность в ответе и объясни, почему ты уверен или не уверен Задача: [здесь ваша математическая задача] [======]

Объяснение эффективности

Данный промпт использует знания из исследования следующим образом:

Учитывает проблему с неопределенностью в задачах рассуждения - мы явно указываем модели на эту проблему **Применяет пошаговое рассуждение** - исследование предлагает "сначала выполнить несколько шагов рассуждения" перед оценкой неопределенности **Запрашивает явную оценку уверенности** - заставляет модель рефлексировать над каждым шагом **Просит альтернативные подходы** - это помогает снизить вероятность ошибки через диверсификацию методов решения Такой промпт помогает компенсировать естественную слабость языковых моделей в определении собственной неуверенности для задач рассуждения, о которой говорится в исследовании.