Доверься мне, я ошибаюсь: Гиперточные галлюцинации в больших языковых моделях

Дата: 2025-02-18 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2502.12964

Рейтинг: 75
Адаптивность: 85

Ключевые выводы:

Исследование направлено на изучение феномена высокоуверенных галлюцинаций в больших языковых моделях (LLM). Основной вывод: LLM могут генерировать галлюцинации с высокой степенью уверенности даже когда обладают правильными знаниями. Это явление, названное CHOKE (Certain Hallucinations Overriding Known Evidence), существует во всех исследованных моделях и не может быть объяснено простым шумом.

Объяснение метода:

Исследование раскрывает критически важный феномен CHOKE - высокоуверенные галлюцинации в LLM даже при наличии правильного знания. Это фундаментально меняет представление о надежности моделей и предлагает практичный метод проверки ответов через переформулировку вопросов. Результаты применимы всеми пользователями без технических знаний, но исследование не дает готовых решений проблемы.

Ключевые аспекты исследования: 1. **Феномен CHOKE (Certain Hallucinations Overriding Known Evidence)** - исследование выявило, что языковые модели могут генерировать галлюцинации с высокой уверенностью даже тогда, когда они обладают правильным знанием. Это противоречит распространенному предположению, что галлюцинации связаны с неуверенностью модели.

Методология обнаружения - авторы разработали трехэтапный подход: (1) выявление примеров, где модель знает правильный ответ, (2) создание вариаций запроса, провоцирующих галлюцинации, и (3) измерение уверенности модели в галлюцинациях с помощью трех метрик (вероятность, разница вероятностей, семантическая энтропия).

Устойчивость феномена - CHOKE наблюдается у различных моделей (Mistral, Llama, Gemma), включая инструктированные версии и модели большего размера, что показывает системность проблемы.

Неэффективность существующих методов снижения галлюцинаций - современные подходы, основанные на оценке уверенности модели, оказались неспособны эффективно выявлять и устранять галлюцинации с высокой уверенностью.

Последовательность CHOKE-примеров - исследование показало, что модели склонны генерировать одни и те же галлюцинации с высокой уверенностью в разных контекстах, что подтверждает неслучайную природу феномена.

Дополнение:

Для работы методов данного исследования не требуется дообучение или API. Хотя ученые использовали доступ к вероятностям токенов и другим техническим метрикам для точного измерения уверенности модели, основные концепции и подходы можно адаптировать и применить в стандартном чате.

Ключевые концепции и подходы, применимые в стандартном чате:

Тестирование знаний с вариациями запросов: Пользователь может сначала задать прямой вопрос, чтобы проверить, знает ли модель ответ Затем переформулировать тот же вопрос в другом контексте (например, используя "детскую" формулировку или вставляя вопрос в диалог) Сравнить ответы на оба запроса для выявления несоответствий

Проверка согласованности:

Задавать один и тот же вопрос несколько раз с небольшими вариациями Если ответы существенно различаются, это может указывать на CHOKE

Запрос самооценки уверенности:

Просить модель оценить свою уверенность в ответе Сравнивать эти самооценки с фактической точностью ответов

Множественные перепроверки:

Для важной информации запрашивать модель объяснить ответ разными способами Проверять внутреннюю согласованность объяснений Ожидаемые результаты: - Выявление противоречий в ответах модели на один и тот же вопрос в разных контекстах - Понимание, в каких областях модель склонна к уверенным галлюцинациям - Повышение общей надежности получаемой от модели информации за счет использования нескольких подходов к проверке - Возможность отличить случаи, когда модель действительно не знает ответа, от случаев CHOKE

Анализ практической применимости: **Феномен CHOKE (высокоуверенные галлюцинации)** - Прямая применимость: Пользователи должны знать, что высокая уверенность LLM не гарантирует точность ответа. Это важно для критической

оценки ответов в серьезных задачах. - Концептуальная ценность: Понимание того, что модель может быть уверена в неправильном ответе даже когда "знает" правильный, меняет подход к оценке надежности LLM. - Потенциал для адаптации: Пользователи могут разработать стратегии дополнительной проверки информации, особенно когда ответы кажутся неожиданными, независимо от уверенности модели.

Методология измерения уверенности модели - Прямая применимость: Ограниченная, так как требует технических знаний и доступа к вероятностям токенов. - Концептуальная ценность: Высокая, демонстрирует разные способы оценки уверенности модели. - Потенциал для адаптации: Пользователи могут адаптировать идею запроса модели о её уверенности в ответе или использовать множественные переформулировки вопроса.

Тестирование с вариациями запросов - Прямая применимость: Пользователи могут переформулировать важные вопросы разными способами, чтобы проверить согласованность ответов. - Концептуальная ценность: Высокая, показывает, как незначительные изменения в формулировке могут вызвать галлюцинации. - Потенциал для адаптации: Легко применимо в повседневном использовании LLM для повышения надежности.

Ограничения существующих методов снижения галлюцинаций - Прямая применимость: Пользователи должны понимать, что даже когда модель отказывается от ответа из-за неуверенности, это не устраняет все галлюцинации. - Концептуальная ценность: Высокая, подчеркивает необходимость дополнительных методов проверки. - Потенциал для адаптации: Пользователи могут комбинировать несколько подходов для проверки ответов.

Prompt:

Использование знаний о галлюцинациях LLM в промптах ## Ключевые выводы исследования для создания промптов

Исследование "Доверься мне, я ошибаюсь" показывает, что языковые модели могут генерировать высокоуверенные галлюцинации (феномен CHOKE), даже когда обладают правильными знаниями. Инструктированные модели демонстрируют ещё худшую калибровку между уверенностью и точностью.

Пример промпта с учетом этих знаний

[======] Я хочу получить фактически точную информацию о [тема]. Учитывая, что даже при высокой уверенности языковые модели могут галлюцинировать:

Предоставь мне ответ на вопрос: [конкретный вопрос]

Для каждого фактического утверждения в своем ответе:

Укажи степень уверенности (высокая/средняя/низкая) Отметь, какие утверждения

могут требовать дополнительной проверки Приведи альтернативные формулировки для проверки согласованности информации

Предложи 2-3 перефразированных варианта моего исходного вопроса, которые могли бы выявить возможные несоответствия в ответе.

Если тебе не хватает информации или ты не уверен, четко обозначь это вместо предположений. [======]

Как это работает

Учет феномена CHOKE: Промпт признает возможность высокоуверенных галлюцинаций и требует явной оценки уверенности для каждого утверждения.

Перефразирование запросов: Исследование показало, что галлюцинации могут быть контекстно-зависимыми, поэтому запрос на альтернативные формулировки помогает выявить несоответствия.

Множественные проверки: Запрос на альтернативные формулировки вопроса помогает обойти контекстную зависимость галлюцинаций.

Признание неопределенности: Явное разрешение модели признавать неуверенность снижает риск генерации "уверенных" но неточных ответов.

Этот подход не устраняет полностью риск галлюцинаций, но создает многоуровневую систему проверки, делая их более заметными для пользователя.