# Предсказание производительности черных ящиков LLM через самозапросы

Дата: 2025-02-16 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2501.01558

Рейтинг: 70
Адаптивность: 85

## Ключевые выводы:

Исследование направлено на разработку метода предсказания производительности LLM в режиме черного ящика (без доступа к внутренним представлениям модели). Основной результат - метод QueRE, который использует дополнительные запросы к модели для извлечения информативных признаков, позволяющих точно предсказывать корректность ответов LLM, обнаруживать модели, подверженные вредоносному влиянию, и различать разные архитектуры моделей.

## Объяснение метода:

Исследование предлагает метод QueRE, позволяющий через дополнительные вопросы оценивать надежность ответов LLM в режиме черного ящика. Высокая ценность для пользователей в оценке достоверности информации и выявлении потенциально неверных ответов. Метод не требует технических знаний для базового применения, но полный потенциал раскрывается при технической реализации. Концепция легко адаптируема для различных сценариев использования.

## Ключевые аспекты исследования: 1. **Метод извлечения черт модели через вопросы**: Исследование предлагает метод QueRE (Question Representation Elicitation), позволяющий извлекать информативные признаки из LLM в режиме "черного ящика" через серию дополнительных вопросов о сгенерированном ответе.

**Предсказание эффективности LLM**: Авторы демонстрируют, что линейные модели, обученные на признаках QueRE, способны надежно предсказывать корректность ответов LLM, часто превосходя методы, требующие доступа к внутренним представлениям модели.

**Обнаружение моделей с вредоносным поведением**: Метод позволяет выявлять случаи, когда LLM находится под влиянием вредоносных системных промптов, которые заставляют модель отвечать неправильно.

**Различение архитектур и размеров моделей**: QueRE может надежно определять

различия между разными моделями и их размерами, что полезно для проверки, действительно ли через API предоставляется заявленная модель.

**Теоретический анализ и гарантии**: Исследование включает теоретический анализ метода с математическими гарантиями его работоспособности даже при аппроксимации вероятностей через выборку.

## Дополнение: Для применения методов этого исследования **не требуется** дообучение модели или специальный API. Хотя авторы использовали API для получения вероятностей токенов для более точных результатов, они также доказали, что метод работает даже при использовании обычного сэмплирования (случайных выборок) из модели.

Концепции и подходы, применимые в стандартном чате:

**Базовый подход проверки уверенности** - задавать модели вопросы типа "Уверены ли вы в своем ответе?", "Можете ли вы объяснить свой ответ?", "Считаете ли вы свой ответ правильным?" после получения основного ответа.

**Множественные уточняющие вопросы** - исследование показывает, что использование разнообразных вопросов (до 40-50 вопросов) дает лучшие результаты, но даже небольшой набор из 5-10 вопросов значительно улучшает оценку надежности.

**Использование случайных последовательностей текста** - интересный вывод исследования заключается в том, что даже отправка модели случайных фраз после основного ответа может выявить информацию о надежности модели.

**Постепенное увеличение количества вопросов** - исследование показывает, что добавление большего числа вопросов улучшает результаты, хотя с убывающей отдачей.

Результаты, которые можно получить: - Более точная оценка достоверности ответов модели - Выявление случаев, когда модель, вероятно, ошибается - Определение того, насколько модель "уверена" в своих ответах - Возможность обнаружить, что модель находится под влиянием вредоносных инструкций

Эти подходы особенно ценны в ситуациях, когда точность информации критически важна, например, в образовании, исследованиях или принятии решений на основе ответов LLM.

## Анализ практической применимости: **Метод извлечения черт модели через вопросы** - Прямая применимость: Высокая. Пользователи могут использовать набор дополнительных вопросов для оценки надежности ответов LLM даже без технических знаний. - Концептуальная ценность: Значительная. Демонстрирует, что дополнительные вопросы модели раскрывают полезную информацию о её внутреннем состоянии. - Потенциал для адаптации: Высокий. Метод не требует доступа к API вероятностей и может работать даже при использовании случайных

выборок из модели.

**Предсказание эффективности LLM** - Прямая применимость: Средняя. Требует обучения линейной модели, что усложняет прямое применение обычными пользователями. - Концептуальная ценность: Высокая. Дает понимание того, как модель "относится" к своим собственным ответам. - Потенциал для адаптации: Высокий. Может быть встроен в пользовательские интерфейсы для предупреждения о потенциально неверных ответах.

**Обнаружение моделей с вредоносным поведением** - Прямая применимость: Средняя. Полезно для разработчиков и организаций, но сложнее для рядовых пользователей. - Концептуальная ценность: Высокая. Демонстрирует возможность обнаружения "отравленных" моделей. - Потенциал для адаптации: Средний. Может быть интегрирован в системы мониторинга, но требует технической реализации.

**Различение архитектур и размеров моделей** - Прямая применимость: Низкая для обычных пользователей, высокая для организаций. - Концептуальная ценность: Средняя. Важно для проверки соответствия заявленным характеристикам. - Потенциал для адаптации: Средний. Полезно для аудита API, но имеет ограниченное применение для рядовых пользователей.

**Теоретический анализ и гарантии** - Прямая применимость: Низкая. Слишком технический аспект для широкой аудитории. - Концептуальная ценность: Средняя. Подтверждает надежность метода, но сложен для понимания. - Потенциал для адаптации: Средний. Обеспечивает уверенность в методе, но требует математических знаний.

## Prompt:

Применение исследования QueRE в промптах для GPT ## Ключевые идеи исследования для промптов

Исследование QueRE показывает, что можно предсказывать производительность LLM и выявлять проблемы, задавая моделям дополнительные вопросы о их собственных ответах. Эта техника "самозапросов" позволяет:

Оценивать надежность ответов Обнаруживать вредоносные влияния Различать архитектуры моделей ## Пример промпта с применением техники QueRE

[======] Я хочу, чтобы ты ответил на мой вопрос, а затем провел самоанализ своего ответа, используя технику QueRE.

ВОПРОС: [Ваш основной вопрос, например о сложной научной концепции]

После того, как ты дашь ответ, пожалуйста, ответь на следующие вопросы о своем ответе: 1. Насколько ты уверен в точности своего ответа по шкале от 1 до 10? 2. Какие части твоего ответа наиболее подвержены ошибкам? 3. Какие источники или

знания ты использовал для формирования ответа? 4. Есть ли альтернативные точки зрения, которые ты не включил? 5. Как бы ты улучшил свой ответ при наличии дополнительной информации?

Используй эти самозапросы, чтобы оценить качество своего ответа и указать на возможные ограничения. [======]

## Как это работает

Этот промпт использует ключевой принцип исследования QueRE — извлечение метаинформации через дополнительные запросы после основного ответа. Когда модель вынуждена анализировать собственный ответ, она:

• Выявляет области неопределенности (калибровка уверенности)

• Указывает на потенциальные слабые места в рассуждении

• Предоставляет контекст о своих источниках знаний

• Демонстрирует осведомленность о возможных ограничениях

Это позволяет пользователю лучше оценить надежность полученной информации без необходимости доступа к внутренним представлениям модели.

## Дополнительные применения

• Для критически важных задач: Включите самозапросы для оценки надежности ответов

• Для обнаружения предвзятости: Попросите модель оценить, не содержит ли ответ предвзятых суждений

• Для сложных решений: Используйте самозапросы для получения более полного понимания уверенности модели

Техника QueRE особенно полезна, когда важна точность и надежность ответов GPT в сценариях с высокой ответственностью.