О правдивости 'удивительно вероятных' ответов больших языковых моделей

Дата: 2025-01-25 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2311.07692

Рейтинг: 75
Адаптивность: 85

Ключевые выводы:

Исследование направлено на повышение фактической точности ответов больших языковых моделей (LLM) с помощью принципа 'удивительно вероятных' ответов. Основной результат показывает, что выбор ответов с высоким соотношением условной вероятности ответа к его априорной вероятности значительно повышает точность LLM на различных бенчмарках, особенно на TruthfulQA, где наблюдалось улучшение до 24 процентных пунктов в общей точности и до 70 процентных пунктов в отдельных категориях вопросов.

Объяснение метода:

Исследование предлагает практически применимый метод повышения фактической точности LLM через концепцию "удивительно вероятных" ответов. Подход может быть адаптирован как в виде промптов для обычных пользователей, так и внедрен разработчиками в интерфейсы. Особенно эффективен для противодействия распространенным заблуждениям, с доказанным улучшением точности до 24% в среднем и до 70% в отдельных категориях.

Ключевые аспекты исследования: 1. **Концепция "удивительно вероятных" ответов**: Исследование вводит понятие "удивительно вероятных" текстовых ответов языковых моделей, вдохновленное принципом "удивительно общих" ответов из теории механизмов выявления истинной информации в коллективном разуме.

Математическая формулировка: Авторы предлагают формулу τ(r,q) = P(r|q) / P(r|'?'), которая оценивает ответы не по абсолютной вероятности, а по отношению условной вероятности ответа к его априорной вероятности.

Эмпирическая эффективность: На бенчмарке TruthfulQA метод показал значительное улучшение точности (до 24 процентных пунктов) по сравнению со стандартными подходами, особенно в вопросах, где модели склонны воспроизводить распространенные заблуждения.

Категориальный анализ: Исследование включает детальный анализ

эффективности метода по различным категориям вопросов, выявляя области особенно высокой эффективности (до 70 процентных пунктов улучшения) и ограничения подхода.

Последовательность результатов: Метод показал стабильное улучшение точности не только на TruthfulQA, но и на других бенчмарках (COPA, StoryCloze), демонстрируя универсальность подхода.

Дополнение:

Применимость в стандартном чате без дообучения или API

Исследование не требует дообучения модели или специального API для применения его основных концепций. Метод "удивительно вероятных" ответов может быть адаптирован для использования в стандартном чате следующими способами:

Через многоэтапные запросы: Пользователь может запросить модель дать несколько возможных ответов на вопрос Затем попросить модель оценить, какие из этих ответов могут быть "удивительно вероятными" (необычно точными относительно их популярности) Выбрать ответ, который модель считает "удивительно вероятным"

Через метапромпты:

"Дай ответ на этот вопрос, но учти, что распространенное мнение может быть неверным" "Какой ответ на этот вопрос был бы не самым очевидным, но наиболее точным?" "Прежде чем ответить, подумай: не является ли очевидный ответ распространенным заблуждением?"

Через имитацию процедуры вычисления:

Пользователь может попросить модель сначала дать ответ на общий вопрос (аналог P(r|'?')) Затем дать ответ на конкретный вопрос (аналог P(r|q)) И наконец, сравнить эти два ответа и определить, какие ответы необычно вероятны в контексте вопроса по сравнению с их общей вероятностью Основные результаты, которые можно получить от применения этих концепций: - Значительное повышение точности в вопросах, где существуют распространенные заблуждения - Более информативные ответы в областях, где "очевидные" ответы часто неточны - Улучшение критического мышления модели через принуждение сравнивать "очевидные" и "неочевидные, но точные" ответы

Данные концепции особенно эффективны для фактологических вопросов в таких областях как история, наука, религия, мифы, суеверия и стереотипы, где исследование показало наибольшие улучшения.

Анализ практической применимости: **Концепция "удивительно вероятных" ответов**: - Прямая применимость: Высокая. Пользователи могут модифицировать

запросы к LLM, включая просьбу предоставить не самый вероятный ответ, а тот, который "удивительно вероятен" (например, "Дай ответ, который не просто наиболее вероятен, а необычно вероятен относительно общей частоты"). - Концептуальная ценность: Значительная. Понимание, что LLM может генерировать более точные ответы при использовании отношения вероятностей, а не просто максимальной вероятности, помогает пользователям лучше понимать механизмы генерации ответов. - Потенциал для адаптации: Высокий. Концепцию можно применять в обычных запросах, прося модель "проверить, не является ли очевидный ответ распространенным заблуждением".

Математическая формулировка: - Прямая применимость: Средняя. Обычные пользователи не могут напрямую использовать формулу, но разработчики могут внедрить этот метод в интерфейсы чатов. - Концептуальная ценность: Высокая. Формула демонстрирует, как можно математически определить "удивительность" ответа, что дает представление о работе LLM. - Потенциал для адаптации: Средний. Пользователи могут адаптировать этот принцип, прося модель "сравнить свой ответ с тем, что она бы ответила на общий вопрос".

Эмпирическая эффективность: - Прямая применимость: Высокая. Пользователи могут сразу применять подход с "удивительно вероятными" ответами для вопросов фактологического характера. - Концептуальная ценность: Значительная. Результаты показывают, что LLM хранят более точную информацию, чем может показаться при стандартном использовании. - Потенциал для адаптации: Высокий. Можно разработать простые промпты, реализующие эту идею.

Категориальный анализ: - Прямая применимость: Средняя. Пользователи могут применять метод избирательно, зная в каких категориях вопросов он наиболее эффективен. - Концептуальная ценность: Высокая. Анализ показывает, что модели по-разному реагируют на различные типы вопросов. - Потенциал для адаптации: Средний. Пользователи могут адаптировать свои запросы с учетом категории вопроса.

Последовательность результатов: - Прямая применимость: Высокая. Универсальность подхода позволяет применять его в разных контекстах. - Концептуальная ценность: Значительная. Подтверждает, что метод не является специфичным для одного типа задач. - Потенциал для адаптации: Высокий. Метод можно адаптировать для широкого спектра задач.

Prompt:

Применение исследования "Удивительно вероятных ответов" в промптах для GPT ## Краткое объяснение принципа

Исследование показывает, что ответы, которые **удивительно вероятны** (имеют высокое отношение условной вероятности к априорной), часто более точны, чем просто наиболее вероятные ответы. Это особенно полезно для противодействия распространенным заблуждениям.

Пример промпта для повышения фактической точности

[======] Я хочу получить максимально точный ответ на мой вопрос. Пожалуйста, следуй этому процессу:

Сначала рассмотри мой вопрос и подумай о наиболее распространенных или очевидных ответах, которые могут прийти на ум. Затем подумай о менее очевидных ответах, которые могут быть верными, даже если они кажутся менее вероятными на первый взгляд. Оцени каждый ответ с точки зрения: Насколько он вероятен в контексте моего конкретного вопроса Насколько он вероятен в общем, без контекста вопроса Выбери ответ, который особенно вероятен в контексте вопроса, но не обязательно самый вероятный в общем смысле. Объясни свой выбор и почему ты считаешь его наиболее точным. Мой вопрос: [Какой запретный плод съели Адам и Ева в Эдемском саду?] [======]

Как это работает

Этот промпт имитирует принцип MaxRatio из исследования, побуждая модель:

Рассмотреть популярные ответы (высокая априорная вероятность) **Найти альтернативные ответы** (которые могут иметь низкую априорную вероятность) **Сравнить условную и априорную вероятности** (хотя модель делает это интуитивно) **Выбрать "удивительно вероятный" ответ** (высокое отношение условной к априорной) Такой подход особенно эффективен для вопросов, где существуют распространенные заблуждения (например, что Адам и Ева съели именно яблоко, хотя в Библии конкретный фрукт не указан).

Другие возможные применения

• Для проверки фактов: "Оцени распространенное утверждение X и менее известные альтернативные объяснения"

• Для многовариантных вопросов: "Рассмотри каждый вариант не только с точки зрения его вероятности в контексте вопроса, но и его общей вероятности"

• Для исторических событий: "Рассмотри как общепринятые, так и менее известные, но потенциально более точные интерпретации"