Кулинарная книга чисел: Понимание чисел в языковых моделях и способы его улучшения

Дата: 2025-03-05 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2411.03766

Рейтинг: 68
Адаптивность: 75

Ключевые выводы:

Исследование направлено на комплексное изучение способностей языковых моделей (LLM) к пониманию и обработке чисел (NUPA). Основной вывод: современные LLM, несмотря на впечатляющие способности к рассуждению, часто допускают ошибки в базовых числовых операциях, особенно при увеличении длины чисел или использовании нестандартных числовых представлений.

Объяснение метода:

Исследование дает ценное понимание ограничений LLM в числовых вычислениях и предлагает стратегии улучшения через формулировку запросов и chain-of-thought. Пользователи могут применять эти знания для повышения точности, особенно разбивая сложные операции на шаги и проверяя результаты. Ограничена доступность технических методов улучшения для обычных пользователей.

Ключевые аспекты исследования: 1. **Комплексное тестирование числового понимания**: Авторы создали тест NUPA (Numerical Understanding and Processing Ability), охватывающий 4 числовых представления (целые числа, дроби, числа с плавающей точкой, научная нотация) и 17 задач в 4 категориях, что дает 41 значимую комбинацию для оценки числовых способностей LLM.

Выявление проблем с числовыми вычислениями: Исследование показало, что даже современные LLM совершают неожиданные ошибки в базовых числовых операциях, особенно при увеличении длины чисел или усложнении задач, несмотря на их способность решать сложные математические задачи.

Анализ методов улучшения: Авторы исследовали три подхода к улучшению NUPA: методы на этапе предобучения (токенизация, позиционное кодирование, форматы данных), дообучение существующих моделей и использование цепочки рассуждений (chain-of-thought).

Оценка токенизации чисел: Исследование показало, что токенизация по одной цифре наиболее эффективна для числовых вычислений, в отличие от

многоцифровых токенизаторов, используемых в современных моделях.

Тестирование дообучения: Простое дообучение на числовых задачах значительно улучшает NUPA для многих, но не всех задач, при этом применение специализированных техник на этапе дообучения может негативно влиять на уже обученные модели.

Дополнение:

Исследование не требует дообучения или API для применения его основных выводов в стандартном чате. Хотя авторы использовали дообучение и специализированные техники для улучшения моделей, основные концепции и подходы доступны обычным пользователям.

Концепции и подходы для стандартного чата:

Chain-of-thought (CoT) - разбиение сложных числовых операций на последовательность простых шагов. Пример: вместо "сколько будет 345 × 27?" спросить "Давай решим 345 × 27 шаг за шагом: сначала 345 × 7, затем 345 × 20, и сложим результаты".

Понимание проблемных областей - зная, что LLM хуже справляются с длинными числами, дробями и научной нотацией, пользователи могут переформулировать задачи или запрашивать дополнительную проверку.

Проверка промежуточных результатов - для сложных вычислений просить модель показывать промежуточные шаги и проверять их.

Формат представления чисел - использовать более простые представления (например, целые числа вместо дробей или научной нотации).

Выравнивание цифр - при работе с числовыми операциями явно указывать на выравнивание цифр, например: "Сложи 123,45 и 6,789, выравнивая числа по десятичной точке".

Результаты применения этих концепций: - Повышение точности числовых вычислений - Снижение вероятности ошибок при работе с длинными числами - Более надежные результаты при работе с разными числовыми представлениями - Возможность решать более сложные числовые задачи путем их декомпозиции

Анализ практической применимости: 1. **Комплексное тестирование числового понимания**: - Прямая применимость: Высокая. Пользователи могут использовать знание о слабостях LLM при работе с числами для более точной формулировки запросов, разбивая сложные числовые операции на простые шаги. - Концептуальная ценность: Очень высокая. Понимание ограничений LLM в числовых вычислениях помогает пользователям избегать ситуаций, где модель может дать неверный ответ. - Потенциал для адаптации: Средний. Пользователи могут адаптировать свои запросы, учитывая слабые места моделей в числовых операциях.

Выявление проблем с числовыми вычислениями: Прямая применимость: Высокая. Пользователи могут проверять результаты числовых операций, особенно для длинных чисел или сложных представлений. Концептуальная ценность: Высокая. Понимание, что модели могут совершать ошибки в простых операциях, меняет подход к использованию LLM для числовых задач. Потенциал для адаптации: Высокий. Знание о проблемных областях позволяет пользователям разработать стратегии проверки и верификации.

Анализ методов улучшения:

Прямая применимость: Низкая. Большинство методов требуют изменения архитектуры модели или процесса обучения, что недоступно обычным пользователям. Концептуальная ценность: Средняя. Понимание методов улучшения дает представление о возможном развитии LLM. Потенциал для адаптации: Низкий. Методы улучшения в основном применимы только разработчиками моделей.

Оценка токенизации чисел:

Прямая применимость: Низкая. Пользователи не могут изменить токенизацию используемой модели. Концептуальная ценность: Средняя. Понимание влияния токенизации на точность числовых операций. Потенциал для адаптации: Низкий. Требует изменения архитектуры модели.

Тестирование дообучения и chain-of-thought:

Прямая применимость: Средняя. Пользователи могут применять подход chain-of-thought для улучшения числовых вычислений. Концептуальная ценность: Высокая. Понимание, что разбиение задачи на шаги повышает точность. Потенциал для адаптации: Высокий. Метод chain-of-thought может быть применен в обычном чате.

Prompt:

Применение знаний из исследования NUPA в промптах для GPT ## Ключевые выводы для использования в промптах

Исследование "Кулинарная книга чисел" выявило важные ограничения языковых моделей при работе с числами:

• LLM испытывают трудности с длинными числами (>10 цифр)

• Модели хуже работают с нестандартными числовыми форматами (дроби, научная нотация)

• Методы цепочки рассуждений (CoT) значительно улучшают точность

Пример промпта с применением знаний исследования

[======] # Задача: Расчет ипотечного платежа

Мне нужно рассчитать ежемесячный платеж по ипотеке.

Исходные данные: - Сумма кредита: 3,450,000 рублей - Срок: 20 лет - Годовая процентная ставка: 7.8%

Инструкции для расчета: 1. Переведи годовую ставку в месячную (раздели на 12) 2. Переведи срок кредита в месяцы 3. Используй формулу аннуитетного платежа: P = L × [i × (1 + i)^n] / [(1 + i)^n - 1] где P - ежемесячный платеж, L - сумма кредита, i - месячная процентная ставка, n - количество месяцев

Пожалуйста, выполни расчет пошагово, проговаривая каждое действие. После каждого промежуточного вычисления проверь результат и только потом переходи к следующему шагу. [======]

Почему это работает

Разбиение на простые шаги: Промпт разбивает сложную задачу на элементарные операции, что соответствует выводам исследования о необходимости упрощения числовых операций

Избегание длинных чисел: Используются числа с небольшим количеством цифр (менее 10), что снижает вероятность ошибки

Применение цепочки рассуждений (CoT): Включена инструкция выполнять расчет пошагово, проговаривая каждое действие, что реализует метод CoT

Явное указание формулы: Предоставление конкретной формулы помогает модели следовать четкому алгоритму решения (rule-following CoT)

Проверка промежуточных результатов: Инструкция проверять каждый шаг снижает вероятность накопления ошибок

Такой подход значительно повышает точность числовых вычислений, выполняемых языковыми моделями, согласно результатам исследования NUPA.