Цепочка черновиков: думай быстрее, пиша меньше

Дата: 2025-03-03 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2502.18600

Рейтинг: 85
Адаптивность: 90

Ключевые выводы:

Исследование представляет новую парадигму Chain of Draft (CoD) для работы с LLM, которая позволяет моделям генерировать минималистичные, но информативные промежуточные рассуждения при решении задач. Основной результат: CoD достигает точности, сравнимой с Chain of Thought (CoT), используя всего 7.6% токенов, что значительно снижает стоимость и задержку при сохранении качества ответов.

Объяснение метода:

Chain of Draft - исключительно практичный метод, позволяющий пользователям сократить использование токенов на 80-92% при сохранении точности. Простая инструкция в промпте заставляет LLM генерировать краткие рассуждения вместо многословных. Метод работает на разных задачах и моделях, но имеет ограничения при zero-shot использовании и на малых моделях.

Ключевые аспекты исследования: 1. **Chain of Draft (CoD)** - новая парадигма промптинга, вдохновленная человеческим мышлением, где LLM генерируют минималистичные, но информативные промежуточные рассуждения, используя значительно меньше токенов.

Сокращение вербальности - CoD достигает той же или лучшей точности, что и Chain of Thought (CoT), используя всего 7.6-20% токенов, что значительно снижает стоимость и задержку.

Принцип минимализма - вместо подробных промежуточных шагов CoD поощряет LLM генерировать краткие, содержательные выводы на каждом шаге, подобно тому, как люди делают короткие заметки.

Экспериментальное подтверждение - исследование демонстрирует эффективность CoD на различных задачах: арифметические рассуждения (GSM8k), здравый смысл и символьные рассуждения.

Ограничения метода - снижение эффективности при использовании без few-shot

примеров и на маленьких моделях (менее 3B параметров).

Дополнение: Для работы методов этого исследования не требуется дообучение или специальный API. Chain of Draft (CoD) - это чисто промптинговая техника, которую можно применить в любом стандартном чате с LLM.

Исследователи использовали GPT-4o и Claude 3.5 Sonnet для экспериментов, но не модифицировали сами модели. Весь метод заключается в специальном формулировании инструкции: "Think step by step but only keep a minimum draft for each thinking step, with 5 words at most."

Концепции и подходы, которые можно применить в стандартном чате:

Ограничение слов в промежуточных шагах - можно прямо указать LLM использовать не более 5 слов на каждый шаг рассуждения.

Формат математических выражений - вместо словесных объяснений использовать компактную математическую нотацию (например, "20 - 12 = 8" вместо "Изначально у Джейсона было 20 леденцов, после он отдал часть и осталось 12...").

Минимализм в рассуждениях - общий принцип "писать меньше, думать больше" применим к любым задачам, требующим рассуждений.

Структура "рассуждение + ответ" - сохранение четкого разделения между рассуждением и финальным ответом (с разделителем ####).

Возможные результаты применения: - Снижение стоимости использования LLM на 80-92% - Сокращение времени ожидания ответа на 48-76% - Более структурированные и легко отслеживаемые рассуждения - Возможность решать более сложные задачи в рамках контекстного окна

Метод особенно эффективен для задач, требующих многошаговых рассуждений: математические задачи, логические головоломки, анализ данных, планирование и другие сценарии, где важен процесс рассуждения, а не только конечный ответ.

Анализ практической применимости: 1. **Chain of Draft (CoD)** - Прямая применимость: Очень высокая. Пользователи могут немедленно применить этот метод, просто добавив в промпт инструкцию "Think step by step but only keep a minimum draft for each thinking step, with 5 words at most". - Концептуальная ценность: Высокая. Метод демонстрирует, что эффективные рассуждения не требуют многословности, что помогает лучше понять работу LLM. - Потенциал для адаптации: Высокий. Принцип краткости может быть применен к различным типам задач и другим методам рассуждений.

Сокращение вербальности Прямая применимость: Высокая. Пользователи могут значительно сократить расходы на токены и время ответа. Концептуальная ценность: Средняя. Понимание соотношения краткости и точности помогает оптимизировать взаимодействие с LLM. Потенциал для адаптации: Высокий.

Принцип минимализма применим к различным контекстам и задачам.

Принцип минимализма

Прямая применимость: Высокая. Легко реализуемый подход для повседневного использования. Концептуальная ценность: Высокая. Помогает понять, что LLM могут эффективно работать с абстрактными, сжатыми промежуточными представлениями. Потенциал для адаптации: Высокий. Может быть адаптирован для различных типов задач и комбинирован с другими методами.

Экспериментальное подтверждение

Прямая применимость: Средняя. Результаты показывают, для каких задач метод наиболее эффективен. Концептуальная ценность: Высокая. Демонстрирует эффективность метода на различных типах задач. Потенциал для адаптации: Средний. Результаты могут быть использованы для выбора оптимальной стратегии в зависимости от задачи.

Ограничения метода

Прямая применимость: Высокая. Знание ограничений помогает пользователям выбрать правильный подход. Концептуальная ценность: Высокая. Понимание зависимости от few-shot примеров и размера модели важно для эффективного использования. Потенциал для адаптации: Средний. Ограничения могут быть преодолены через адаптацию метода или комбинирование с другими подходами.

Prompt:

Использование Chain of Draft (CoD) в промптах для GPT ## Ключевая идея исследования Chain of Draft (CoD) - это подход, позволяющий моделям генерировать минималистичные, но информативные промежуточные рассуждения. Он достигает точности, сравнимой с Chain of Thought (CoT), но использует всего ~8% токенов, что снижает стоимость и задержку.

Пример промпта с применением CoD

[======] Реши следующую математическую задачу, используя подход Chain of Draft. Для каждого шага рассуждения используй не более 5 слов.

Задача: В магазине продаются футболки по 800 рублей и джинсы по 2400 рублей. Маша купила 3 футболки и 2 пары джинсов. Сколько всего денег она потратила?

Инструкция: Думай пошагово, но записывай только краткие черновые мысли для каждого шага, максимум 5 слов на шаг. После рассуждений дай финальный ответ. [======]

Как это работает

Минимизация токенов: Вместо развернутых рассуждений (как в CoT) модель генерирует очень краткие промежуточные мысли, что сокращает количество используемых токенов на ~92%.

Сохранение точности: Несмотря на краткость, такой подход позволяет моделям сохранять или даже улучшать точность ответов по сравнению с полными рассуждениями.

Снижение задержки: Сокращение объема генерируемого текста значительно уменьшает время ожидания ответа (на 48-76% по данным исследования).

Экономия ресурсов: Меньшее количество токенов = меньшая стоимость использования API и меньшая вычислительная нагрузка.

Практическое применение

Этот подход особенно полезен для: - Приложений, работающих в реальном времени - Ситуаций с ограниченным бюджетом на API - Мобильных приложений с ограниченными ресурсами - Случаев, когда важна скорость получения ответа

Вы можете адаптировать инструкцию "Think step by step but only keep a minimum draft for each thinking step, with 5 words at most" для различных типов задач, требующих рассуждений.