Скамейка LCTG: Бенчмарк генерации текста с контролем LLM

Дата: 2025-01-27 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2501.15875

Рейтинг: 75
Адаптивность: 80

Ключевые выводы:

Исследование представляет LCTG Bench - первый японский бенчмарк для оценки контролируемости (управляемости) больших языковых моделей (LLM) при генерации текста. Основная цель - создать унифицированную систему оценки способности LLM следовать конкретным инструкциям при генерации текста на японском языке. Результаты показали значительный разрыв в производительности между многоязычными моделями (GPT-4, GPT-3.5, Gemini-Pro) и японскими моделями, а также выявили общие проблемы с контролем количества символов во всех моделях.

Объяснение метода:

Исследование предлагает универсальную методологию контроля генерации текста по четырем аспектам (формат, количество символов, ключевые/запрещенные слова), применимую в любых LLM. Представленные структуры промптов и подходы к оценке могут быть непосредственно использованы пользователями для повышения качества взаимодействия с чат-моделями. Выявленные особенности разных моделей помогают выбрать оптимальную для конкретных задач.

Ключевые аспекты исследования: 1. **LCTG Bench** - первый японский бенчмарк для оценки управляемости (контролируемости) LLM при генерации текста, позволяющий выбрать наиболее подходящую модель для различных сценариев использования.

Четыре аспекта контролируемости генерации текста: Format (формат), Character Count (количество символов), Keyword (ключевые слова), Prohibited Word (запрещенные слова), которые оцениваются единообразно в трех задачах.

Три генеративные задачи: Summarization (суммаризация), Ad Text Generation (генерация рекламного текста) и Pros & Cons Generation (генерация плюсов и минусов), каждая с различными характеристиками для всесторонней оценки.

Методология оценки: использование правило-ориентированной проверки для измерения контролируемости и GPT-4 для оценки качества генерируемого

содержимого.

Выявление разрыва в производительности между многоязычными моделями (GPT-4, GPT-3.5, Gemini-Pro) и японскими моделями в контексте контролируемости генерации текста.

Дополнение: Для работы с методами этого исследования не требуется дообучение или API. Хотя авторы использовали GPT-4 для оценки качества и постобработки результатов, основные концепции и подходы полностью применимы в стандартном чате с любой LLM.

Концепции и подходы, которые можно применить в стандартном чате:

Четыре аспекта контролируемости: FORMAT: указание в промпте "выведи только результат, без пояснений" C-COUNT: указание точного количества символов/слов в выводе KEYWORD: требование использовать определенные ключевые слова P-WORD: запрет на использование определенных слов

Структура промптов:

Трехчастная структура: инструкция задачи + условие + базовый текст Четкое разделение условий от основной инструкции

Понимание ограничений моделей:

Учет того, что контроль количества символов может быть проблематичным Подготовка к тому, что модель может добавлять объяснения, даже если их не просили Применяя эти концепции, пользователи могут получить: - Более точное соответствие выводов заданным требованиям - Лучшее понимание ограничений моделей - Более эффективные стратегии формулирования запросов - Возможность контролировать включение/исключение определенного содержимого

Примечательно, что даже GPT-4 показывает низкие результаты при контроле точного количества символов, что подсказывает пользователям необходимость проверки и возможной постобработки результатов при работе с такими ограничениями.

Анализ практической применимости: **1. Четыре аспекта контролируемости** - **Прямая применимость**: Пользователи могут сразу использовать эти аспекты как шаблоны для своих промптов при работе с LLM, чтобы контролировать формат, длину, включение или исключение определенных слов. - **Концептуальная ценность**: Понимание того, какие типы ограничений лучше всего работают с LLM и как формулировать их в промптах. - **Потенциал для адаптации**: Эти четыре аспекта применимы к любым языкам и задачам, что делает их универсальным инструментом для пользователей.

2. Методология проверки контролируемости - **Прямая применимость**: Пользователи могут адаптировать подход исследования для проверки соблюдения

условий своих запросов. - **Концептуальная ценность**: Понимание того, как оценивать соответствие выходных данных LLM заданным условиям. - **Потенциал для адаптации**: Метод удаления лишних пояснительных текстов с помощью GPT-4 может быть применен пользователями для постобработки результатов.

3. Выявленные разрывы в производительности моделей - **Прямая применимость**: Пользователи могут использовать результаты для выбора наиболее подходящей модели для своих задач. - **Концептуальная ценность**: Понимание сильных и слабых сторон разных моделей при работе с разными типами ограничений. - **Потенциал для адаптации**: Информация о том, что многоязычные модели лучше справляются с контролируемостью, может помочь пользователям при выборе модели.

4. Примеры промптов для контролируемой генерации - **Прямая применимость**: Пользователи могут напрямую использовать структуру представленных промптов для своих задач. - **Концептуальная ценность**: Понимание, как формулировать условия в промптах для достижения желаемого результата. - **Потенциал для адаптации**: Шаблоны промптов могут быть адаптированы для различных задач и языков.

5. Выводы о сложности контролируемости Character Count - **Прямая применимость**: Пользователи могут учитывать, что даже современные модели вроде GPT-4 испытывают трудности с точным контролем количества символов. - **Концептуальная ценность**: Понимание ограничений LLM при работе с числовыми ограничениями. - **Потенциал для адаптации**: Пользователи могут разработать дополнительные стратегии для обеспечения соответствия ограничениям по длине текста.

Prompt:

Использование результатов исследования LCTG Bench в промптах для GPT ## Ключевые инсайты для создания эффективных промптов

Исследование LCTG Bench предоставляет ценную информацию о контролируемости языковых моделей, которую можно использовать для оптимизации промптов:

Многоязычные модели превосходят специализированные (особенно для японского языка) **Контроль количества символов** — слабое место всех моделей **Четкие инструкции по формату** значительно улучшают результат **Явное указание ключевых и запрещенных слов** требует особого внимания ## Пример оптимизированного промпта

[======] # Задание: Создание рекламного текста для японского ресторана

Требования к формату: - Структура: заголовок, 3 абзаца основного текста, призыв к действию - Длина: ровно 400 символов (не токенов) - Заголовок выделить жирным шрифтом - Не добавлять пояснения до и после текста

Обязательные элементы: - Ключевые слова для включения: "аутентичный", "свежие ингредиенты", "традиции" - Запрещенные слова: "дешевый", "быстрый", "фастфуд"

Дополнительные инструкции: - После создания текста проверь количество символов и скорректируй до точного соответствия требованию в 400 символов - Убедись, что все ключевые слова включены естественным образом - Подтверди отсутствие всех запрещенных слов

Пожалуйста, создай рекламный текст, строго соблюдая все указанные требования. [======]

Почему это работает

Данный промпт учитывает выводы исследования LCTG Bench следующим образом:

Четкая структура формата — исследование показало, что модели лучше справляются с задачами, когда формат четко определен

Явное указание на проверку количества символов — компенсирует слабое место всех моделей (C-COUNT), заставляя модель дополнительно проверить этот параметр

Выделение ключевых и запрещенных слов в отдельные списки — улучшает понимание модели о том, что должно и не должно быть включено

Запрет на пояснительные тексты — решает проблему, когда модели добавляют ненужные пояснения в начале или конце ответа

Дополнительные инструкции по самопроверке — заставляют модель провести внутреннюю валидацию результата перед выдачей ответа

Такой подход к составлению промптов значительно повышает вероятность получения текста, соответствующего всем заданным параметрам контролируемости.