# Оценка управляемости подсказок больших языковых моделей

Дата: 2025-02-15 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2411.12405

Рейтинг: 72
Адаптивность: 80

## Ключевые выводы:

Исследование направлено на оценку способности больших языковых моделей (LLM) к управлению через промпты. Основная цель - разработать метрику для измерения того, насколько модель может быть 'настроена' на отражение различных персон и ценностных систем с помощью промптов. Результаты показывают, что текущие модели имеют ограниченную управляемость из-за асимметрии в их базовом поведении и сопротивления изменениям в определенных направлениях.

## Объяснение метода:

Исследование предоставляет ценную методологию для измерения и понимания стерируемости LLM через промпты. Основные выводы о количестве необходимых направляющих утверждений, асимметрии стерируемости и различиях между моделями напрямую применимы к разработке эффективных стратегий промптинга. Требует некоторых технических знаний, но концепции адаптируемы для обычных пользователей.

## Ключевые аспекты исследования: 1. **Формальное определение стерируемости LLM** - исследование вводит методологию для оценки того, насколько модели могут быть "направлены" с помощью промптов для отражения различных персон. Ключевая концепция - это "профиль оценки", представляющий поведение модели при ответе на определенные вопросы.

**Индексы стерируемости** - авторы разработали количественные метрики для измерения степени, в которой модель может быть направлена в определенном направлении с помощью промптов, с учетом базового поведения модели.

**Кривые стерируемости** - визуализация того, как поведение модели меняется при увеличении "бюджета стерирования" (количества направляющих утверждений в промпте).

**Бенчмарк многомерных персон** - эксперименты оценивают стерируемость моделей по 32 измерениям личности, от этических убеждений до личностных черт.

**Асимметричная стерируемость** - исследование выявило, что модели часто легче направить в одном направлении, чем в другом, и имеют предвзятость в сторону определенных измерений.

## Дополнение:

Для работы методов этого исследования не требуется дообучение или специальный API - основные концепции и подходы могут быть адаптированы для использования в стандартном чате. Хотя авторы использовали доступ к логарифмическим вероятностям для точного измерения стерируемости, обычные пользователи могут применять ключевые идеи без этого:

**Техника "Принципов"** - Включение направляющих утверждений в начало запроса в формате "Вы придерживаетесь следующих принципов: [принципы]" эффективно влияет на поведение модели.

**Оптимальное количество направляющих утверждений** - Исследование показывает, что часто достаточно 1-3 направляющих утверждения, после чего эффект насыщается, особенно для более продвинутых моделей.

**Асимметрия стерируемости** - Понимание того, что модели легче направить в сторону определенных значений (например, в сторону этичности и вежливости), может помочь пользователям сформулировать более эффективные запросы.

**Измерения для направления** - Пользователи могут фокусироваться на конкретных измерениях личности (открытость, добросовестность, экстраверсия и т.д.) при направлении модели.

Применяя эти концепции, пользователи могут получить более персонализированные, последовательные и предсказуемые ответы от LLM в стандартном чате без необходимости в специальных инструментах или API.

## Анализ практической применимости: **Формальное определение стерируемости LLM**: - Прямая применимость: Пользователи могут адаптировать методологию для тестирования пределов возможностей LLM принимать различные точки зрения. - Концептуальная ценность: Высокая; обеспечивает структурированное понимание того, как модели реагируют на направляющие подсказки. - Потенциал для адаптации: Пользователи могут использовать общий подход для разработки собственных тестов направляемости в конкретных областях.

**Индексы стерируемости**: - Прямая применимость: Средняя; требует технических знаний для реализации, но дает объективные метрики для оценки эффективности направления модели. - Концептуальная ценность: Помогает понять пределы изменения поведения LLM через промпты. - Потенциал для адаптации: Метрики могут быть адаптированы для измерения эффективности различных техник формирования промптов.

**Кривые стерируемости**: - Прямая применимость: Пользователи могут определить оптимальное количество направляющих утверждений для эффективного изменения поведения модели. - Концептуальная ценность: Демонстрирует, что большие/более совершенные модели часто требуют меньше направляющих утверждений. - Потенциал для адаптации: Высокий; подход можно использовать для оптимизации промптов в практических задачах.

**Бенчмарк многомерных персон**: - Прямая применимость: Пользователи могут определить, по каким измерениям конкретная модель лучше поддается направлению. - Концептуальная ценность: Выявляет врожденные предпочтения и ограничения моделей. - Потенциал для адаптации: Методология может быть применена к другим аспектам поведения LLM.

**Асимметричная стерируемость**: - Прямая применимость: Пользователи могут учитывать эти асимметрии при формулировании промптов. - Концептуальная ценность: Высокая; показывает, что базовое поведение модели сильно влияет на возможность ее направления. - Потенциал для адаптации: Знание об асимметриях может помочь в разработке более эффективных стратегий промптинга.

## Prompt:

Применение исследования управляемости LLM в промптах ## Ключевые выводы для использования

Исследование показывает, что: - Модели имеют асимметричную управляемость (легче "направлять" в отрицательном направлении) - Наиболее управляемы измерения этики/философии и личности - Более продвинутые модели требуют меньше инструкций для управления - У каждой модели есть свое базовое поведение, которое ограничивает диапазон управляемости

## Пример эффективного промпта с применением знаний из исследования

[======] # Промпт для создания этического анализа с консервативным уклоном

## Контекст и инструкции Ты - консервативный этический аналитик с опытом в традиционных ценностях. Я хочу, чтобы ты проанализировал следующую ситуацию с точки зрения традиционных ценностей.

## Примеры твоих убеждений (для настройки твоего ответа) - Традиционная семья - основа здорового общества - Личная ответственность важнее коллективной - Постепенные изменения предпочтительнее радикальных реформ - Уважение к устоявшимся институтам и традициям необходимо

## Задание Проанализируй следующую ситуацию: [описание ситуации]

Структурируй свой ответ, включая: 1. Ключевые этические принципы, применимые к

ситуации 2. Анализ с точки зрения традиционных ценностей 3. Рекомендации, основанные на консервативном подходе [======]

## Объяснение эффективности

Данный промпт учитывает результаты исследования следующим образом:

**Фокус на этике** — использует область, где модели наиболее управляемы (этика/философия) **Конкретные примеры убеждений** — предоставляет небольшое, но целенаправленное количество инструкций **Четкое направление** — задает конкретное направление (консервативный уклон), учитывая базовое поведение модели **Структурированность** — помогает модели следовать заданному направлению через четкую структуру ответа Такой подход повышает вероятность того, что модель будет следовать заданной "персоне" и ценностной системе, оптимально используя ее возможности управляемости.