PReasoning о теории разума на основе гипотез для больших языковых моделей

Дата: 2025-02-17 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2502.11881

Рейтинг: 78
Адаптивность: 85

Ключевые выводы:

Исследование представляет алгоритм Thought Tracing для улучшения способности больших языковых моделей (LLM) отслеживать и выводить ментальные состояния агентов в тексте. Основная цель - разработать метод, который может отслеживать мысли и убеждения персонажей в тексте без опоры на заранее известные ответы. Результаты показывают, что этот алгоритм значительно улучшает производительность LLM на задачах теории сознания (Theory of Mind), превосходя базовые модели и специализированные модели рассуждения.

Объяснение метода:

Исследование представляет высокую ценность, предлагая метод улучшения взаимодействия с LLM в задачах понимания намерений. Алгоритм Thought Tracing дает практический подход к структурированию запросов, демонстрирует способы преодоления ограничений моделей и работы с неопределенностью. Основные концепции доступны для адаптации, хотя полная реализация требует технических знаний.

Ключевые аспекты исследования: 1. **Алгоритм Thought Tracing** - новый метод для отслеживания и вывода ментальных состояний агентов в тексте, основанный на принципах байесовской теории разума (BToM) и алгоритме Sequential Monte Carlo. Алгоритм генерирует множество гипотез о мыслях агента и взвешивает их на основе наблюдений.

Естественно-языковое представление гипотез - в отличие от традиционных вероятностных моделей, гипотезы о ментальных состояниях представлены в виде естественного языка и генерируются языковыми моделями.

Улучшение производительности моделей в задачах Theory of Mind - алгоритм значительно улучшает способность языковых моделей отвечать на вопросы, связанные с пониманием намерений и ментальных состояний агентов, без специального дообучения.

Сравнение с моделями рассуждения - исследование выявило, что модели, специализирующиеся на рассуждениях (O1, R1), не демонстрируют такого же превосходства в задачах Theory of Mind, как в математических задачах.

Эффективность в условиях неопределенности - алгоритм специально разработан для работы в социальной сфере, где отсутствуют объективно проверяемые ответы, в отличие от математических или программистских задач.

Дополнение: Исследование представляет алгоритм "Thought Tracing", который действительно можно адаптировать для использования в стандартном чате с LLM без необходимости дообучения или специальных API.

Хотя авторы использовали API для подсчета весов гипотез, они отмечают, что вместо этого можно использовать простой подход с инструкциями для модели выбрать из шести вариантов вероятности (от "очень вероятно" до "очень маловероятно"), что работает даже лучше.

Основные концепции, которые можно применить в стандартном чате:

Разделение текста на состояния и действия - можно попросить модель проанализировать текст, выделив состояния и действия агента.

Генерация гипотез о ментальных состояниях - можно попросить модель сгенерировать несколько (3-4) гипотез о том, что агент мог думать в определенный момент.

Оценка вероятности гипотез - можно попросить модель оценить, насколько вероятно каждое действие агента, учитывая каждую гипотезу.

Обновление гипотез - на основе новых действий можно попросить модель обновить гипотезы.

Суммирование гипотез - в конце можно попросить модель обобщить наиболее вероятные мысли агента.

Пример применения в стандартном чате:

Пользователь: Проанализируй этот текст: "Джон искал по всему дому ключи. Он проверил кухню, гостиную и спальню, но не заглянул в ванную. Затем он вышел из дома."

Выдели состояния и действия Джона. Предложи 3 гипотезы о том, что Джон мог думать после проверки спальни. Оцени, насколько вероятно его действие "выйти из дома" при каждой гипотезе. Какое наиболее вероятное ментальное состояние Джона когда он выходил из дома? Такой подход может значительно улучшить понимание намерений персонажей в текстах, анализ литературных произведений, и даже помочь в интерпретации реальных ситуаций, новостей или поведения людей. Это

особенно ценно для писателей, психологов, аналитиков и всех, кто работает с анализом поведения и намерений.

Анализ практической применимости: 1. **Алгоритм Thought Tracing** - Прямая применимость: Высокая. Пользователи могут адаптировать этот подход для лучшего взаимодействия с ИИ-агентами, прогнозируя их поведение и формулируя запросы, учитывающие ограничения моделей в понимании намерений. - Концептуальная ценность: Очень высокая. Понимание того, что LLM могут генерировать гипотезы о ментальных состояниях и взвешивать их, дает пользователям представление о том, как модели "рассуждают" о социальных ситуациях. - Потенциал для адаптации: Высокий. Метод может быть упрощен для использования в повседневных запросах, например, для анализа текстов, персонажей в литературе или для лучшего понимания мотивации людей в новостях.

Естественно-языковое представление гипотез Прямая применимость: Средняя. Обычные пользователи могут не иметь доступа к внутренним весам модели, но могут запрашивать модель генерировать несколько гипотез о ментальных состояниях персонажей. Концептуальная ценность: Высокая. Понимание того, что LLM могут работать с неопределенностью через множественные гипотезы, помогает пользователям формулировать более эффективные запросы. Потенциал для адаптации: Высокий. Пользователи могут применять принцип множественных гипотез для решения собственных задач анализа намерений и мотивов.

Улучшение производительности моделей

Прямая применимость: Высокая. Пользователи могут структурировать свои запросы, включая в них элементы восприятия и предполагаемых мыслей, что может значительно улучшить качество ответов. Концептуальная ценность: Высокая. Понимание того, что включение промежуточных рассуждений о ментальных состояниях улучшает ответы, дает пользователям инструмент для получения более точных результатов. Потенциал для адаптации: Высокий. Метод может быть адаптирован для использования в различных контекстах анализа текста, психологического анализа и даже для улучшения диалоговых систем.

Сравнение с моделями рассуждения

Прямая применимость: Средняя. Пользователи получают понимание того, что модели, оптимизированные для математических рассуждений, могут не превосходить обычные LLM в социальных задачах. Концептуальная ценность: Высокая. Это помогает пользователям выбирать подходящие модели для разных типов задач и не переоценивать возможности "рассуждающих" моделей. Потенциал для адаптации: Средний. Это знание может помочь пользователям более осознанно выбирать модели для своих задач.

Эффективность в условиях неопределенности

Прямая применимость: Высокая. Пользователи могут применять подходы из исследования для задач, где нет однозначно правильных ответов, например,

анализа литературных персонажей. Концептуальная ценность: Очень высокая. Понимание того, как работать с неопределенностью в социальных контекстах, дает пользователям новые инструменты для работы с LLM. Потенциал для адаптации: Высокий. Принципы работы с неопределенностью могут быть применены к широкому спектру задач вне социального контекста.

Prompt:

Использование Thought Tracing в промтах для GPT ## Суть метода Thought Tracing

Метод Thought Tracing позволяет языковым моделям лучше отслеживать и выводить ментальные состояния персонажей в тексте. Он работает путем: - Генерации множественных гипотез о мыслях персонажей - Взвешивания этих гипотез на основе наблюдаемых действий - Последовательного обновления представлений о ментальных состояниях

Пример промта для анализа литературного произведения

[======] Проанализируй следующий отрывок из романа, используя метод Thought Tracing:

[ТЕКСТ ОТРЫВКА]

Инструкции: 1. Разбей текст на последовательность состояний и действий каждого ключевого персонажа. 2. Для каждого персонажа сгенерируй 3-4 возможные гипотезы о его текущих мыслях, убеждениях и намерениях в каждой ключевой точке повествования. 3. Оцени вероятность каждой гипотезы, основываясь на наблюдаемых действиях персонажа. 4. Для наиболее вероятных гипотез опиши, как они объясняют последующие действия персонажа. 5. В заключении, представь наиболее правдоподобную траекторию мыслей каждого персонажа на протяжении всего отрывка.

Важно: Фокусируйся не только на том, что персонажи знают, но и на том, во что они верят, чего не знают, и как их неполное или ошибочное понимание ситуации влияет на их действия. [======]

Почему это работает

Данный промт использует ключевые аспекты Thought Tracing:

Генерация множественных гипотез - просим модель создать несколько возможных объяснений ментальных состояний **Оценка вероятности** - заставляем модель взвешивать гипотезы на основе действий персонажей **Последовательное обновление** - требуем отслеживать изменения в ментальных состояниях с течением повествования Такой подход позволяет GPT выйти за рамки поверхностного анализа и глубже проникнуть в теорию сознания персонажей, что, согласно исследованию, значительно улучшает качество анализа социальных взаимодействий и понимание мотиваций персонажей.